TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN &

TRUYỀN THÔNG VIỆT - HÀN

**Khoa Khoa Học Máy Tính**



**BÁO CÁO MÔN KHO DỮ LIỆU**

**XÂY DỰNG KHO DỮ LIỆU QUẢN LÝ TAI NẠN GIAO THÔNG TẠI MỸ**

Sinh viên thực hiện:**Phạm Như Quốc Bảo 21IT531**

**Trương Đăng Hoàn 21IT487**

**Nguyễn Thị Mai Lan 21IT558**

**Nguyễn Thị Út Kiều 21IT557**

GVHD: ThS. Trần Thanh Liêm

Đà Nẵng, tháng 10 năm 2024

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG VIỆT - HÀN

**Khoa Khoa Học Máy Tính**



Sinh viên thực hiện:**Phạm Như Quốc Bảo 21IT531**

**Trương Đăng Hoàn 21IT487**

**Nguyễn Thị Mai Lan 21IT558**

**Nguyễn Thị Út Kiều 21IT557**

GVHD: ThS. Trần Thanh Liêm

Đà Nẵng, tháng 10 năm 2024

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

*................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................*

**Chữ kí của giảng viên hướng dẫn**

**LỜI CẢM ƠN**

Kính gửi thầy Trần Thanh Liêm

Em xin chân thành cảm ơn thầy đã tận tình giảng dạy và hướng dẫn em trong suốt học phần kho dữ liệu. Qua sự chỉ dạy của thầy, em không chỉ được trang bị những kiến thức căn bản và chuyên sâu về kho dữ liệu mà còn học hỏi được nhiều kỹ năng quan trọng trong việc phân tích, thiết kế và xây dựng hệ thống kho dữ liệu. Những kiến thức và kỹ năng này chắc chắn sẽ là nền tảng vững chắc cho quá trình học tập và công tác sau này của em.

Em cảm kích trước sự nhiệt huyết và tâm huyết mà thầy đã dành cho học phần này. Sự tận tâm của thầy đã giúp em vượt qua những khó khăn trong việc tiếp thu kiến thức và phát triển tư duy phân tích.

Một lần nữa, em xin chân thành cảm ơn thầy và kính chúc thầy dồi dào sức khỏe, thành công trong sự nghiệp giảng dạy và nghiên cứu, để tiếp tục truyền đạt những kiến thức quý báu đến các thế hệ sinh viên tiếp theo.

Trân trọng cảm ơn!

**MỤC LỤC**

[1 TỔNG QUAN 1](#_2et92p0)

[1.1 Lý do chọn đề tài 1](#_tyjcwt)

[1.2 Mục đích 1](#_1t3h5sf)

[1.3 Phạm vi đề tài 2](#_4d34og8)

[1.4 Đối tượng sử dụng 3](#_2s8eyo1)

[1.5 Phương pháp nghiên cứu 4](#_17dp8vu)

[1.6 Bố cục báo cáo 5](#_3rdcrjn)

[2 CƠ SỞ LÝ THUYẾT 6](#_26in1rg)

[2.1 Visual Studio 6](#_lnxbz9)

[2.2 SSMS 8](#_35nkun2)

[2.3 SSIS 9](#_1ksv4uv)

[2.4 SSAS 10](#_ek1rtv28duxy)

[2.5 SSRS 12](#_44sinio)

[3 PHÂN TÍCH THIẾT KẾ KHO DỮ LIỆU 14](#_2jxsxqh)

[3.1 Mô hình logic 14](#_z337ya)

[3.1.1 Dimensions: 14](#_3j2qqm3)

[3.1.2 Facts: 16](#_1y810tw)

[3.1.3 Sơ đồ mô hình logic: 17](#_4i7ojhp)

[3.2 Mô hình hóa vật lý 17](#_2xcytpi)

[3.2.1 Dimensions: 17](#_1ci93xb)

[3.2.2 Facts: 21](#_3whwml4)

[3.2.3 Sơ đồ mô hình hóa vật lý: 23](#_2bn6wsx)

[4 QUÁ TRÌNH ETL 24](#_qsh70q)

[4.1 Chuẩn bị dữ liệu 24](#_3as4poj)

[4.2 Kết quả 25](#_1pxezwc)

[4.3 Quá trình ETL 25](#)

[4.3.1 Tạo dự án SSIS 25](#)

[4.3.2 Kết nối dữ liệu với sql 26](#)

[4.3.3 Tạo các bảng Dimension 27](#)

[4.3.4 Tạo bảng fact 31](#)

[4.3.5 Xoá dữ liệu của các bảng trước khi đổ dữ liệu 33](#)

[4.3.6 Thêm Khoá ngoại giữa bảng Fact và bảng Dim 34](#)

[1. KẾT QUẢ ĐẠT ĐƯỢC 35](#_49x2ik5)

[2. HƯỚNG PHÁT TRIỂN 35](#_2p2csry)

**DANH MỤC HÌNH**

[Hình 1 Hình Logo Java 11](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.147n2zr)

[Hình 2 Hình Logo MySQL 12](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.3o7alnk)

[Hình 3 hình sơ đồ hệ thống 15](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.23ckvvd)

[Hình 4 Thực thi tệp WebServer.jav 17](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.ihv636)

[Hình 5 Duyệt đến thư mục chứa các tập tin Resource.Zip 18](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.32hioqz)

[Hình 6 Khởi động máy chủ 19](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.1hmsyys)

[*Hình 7 Hình trang chủ* 20](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.41mghml)

[*Hình 8 Hình trang profile* 20](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.2grqrue)

[*Hình 9 Hình trang upload* 21](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.vx1227)

[*Hình 10 Hình trang đăng kí* 22](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.3fwokq0)

[Hình 11 *Hình trang đăng nhập* 23](https://docs.google.com/document/d/19iIGQvURipQ1qsj4TIuj0BY9Z4ZdAaP-/edit#heading=h.1v1yuxt)

# **TỔNG QUAN**

## **Lý do chọn đề tài**

Trong những năm gần đây, tai nạn giao thông trở thành một vấn đề nghiêm trọng tại Mỹ, gây ra nhiều thiệt hại về người và tài sản, cũng như ảnh hưởng lớn đến sự an toàn của người dân và chất lượng cơ sở hạ tầng. Dữ liệu liên quan đến tai nạn giao thông có tính chất phong phú và phức tạp, bao gồm thông tin về thời gian, địa điểm, loại phương tiện, nguyên nhân gây tai nạn và nhiều yếu tố khác. Việc phân tích, xử lý và lưu trữ dữ liệu này có thể hỗ trợ đáng kể cho các cơ quan chức năng trong việc đưa ra các chính sách, biện pháp phòng ngừa và giảm thiểu tai nạn giao thông.

Do đó, đề tài "Xây dựng kho dữ liệu về tai nạn giao thông ở Mỹ" được chọn với mong muốn khai thác tiềm năng của dữ liệu lớn trong lĩnh vực này. Thông qua việc thiết kế và xây dựng kho dữ liệu chuyên dụng, đề tài hướng đến mục tiêu cung cấp một hệ thống lưu trữ và truy xuất dữ liệu có cấu trúc, giúp hỗ trợ quá trình phân tích và đánh giá các xu hướng tai nạn giao thông, từ đó đóng góp cho việc cải thiện an toàn giao thông, tối ưu hóa hạ tầng và nâng cao ý thức tham gia giao thông của cộng đồng.

Ngoài ra, đề tài còn là cơ hội để áp dụng các kiến thức về kho dữ liệu trong thực tế, từ quá trình thu thập, làm sạch, chuyển đổi dữ liệu cho đến thiết kế kiến trúc hệ thống, đảm bảo tính toàn vẹn và khả năng mở rộng của kho dữ liệu. Điều này sẽ giúp củng cố và phát triển năng lực nghiên cứu và ứng dụng các phương pháp khoa học dữ liệu của bản thân trong lĩnh vực phân tích dữ liệu lớn.

## **Mục đích**

Đề tài "Xây dựng kho dữ liệu về tai nạn giao thông ở Mỹ" được thực hiện nhằm đạt được những mục tiêu sau:

**Cung cấp nguồn dữ liệu tập trung và có cấu trúc**: Xây dựng một kho dữ liệu tập trung, dễ dàng truy xuất và quản lý thông tin chi tiết về các vụ tai nạn giao thông ở Mỹ, bao gồm thông tin về thời gian, địa điểm, phương tiện, nguyên nhân và mức độ thiệt hại. Điều này giúp giảm thiểu tình trạng phân tán dữ liệu, đảm bảo tính nhất quán và chính xác khi phân tích dữ liệu.

**Hỗ trợ phân tích và đánh giá xu hướng tai nạn giao thông**: Tạo điều kiện thuận lợi cho việc phân tích sâu hơn về các xu hướng và mẫu hình của tai nạn giao thông. Thông qua các phân tích này, có thể nhận diện được những khu vực nguy hiểm, thời điểm xảy ra tai nạn thường xuyên, và các yếu tố ảnh hưởng chính đến tỷ lệ tai nạn giao thông.

**Hỗ trợ việc xây dựng chính sách và biện pháp an toàn**: Kho dữ liệu cung cấp nền tảng dữ liệu vững chắc cho các cơ quan quản lý và chính phủ để đề xuất các chính sách phòng ngừa, cải thiện cơ sở hạ tầng giao thông và nâng cao ý thức an toàn giao thông. Việc dựa trên dữ liệu cụ thể và khách quan sẽ giúp các chính sách được xây dựng một cách hiệu quả hơn.

**Nâng cao khả năng ứng dụng công nghệ trong quản lý dữ liệu lớn**: Tạo cơ hội ứng dụng các kỹ thuật tiên tiến trong xây dựng và quản lý kho dữ liệu, bao gồm ETL (Extract, Transform, Load), tối ưu hóa truy vấn, và bảo mật dữ liệu. Điều này không chỉ giúp hoàn thiện kỹ năng cá nhân mà còn góp phần ứng dụng công nghệ trong quản lý và phân tích dữ liệu giao thông.

**Cung cấp nền tảng nghiên cứu cho các đề tài liên quan**: Kho dữ liệu có thể phục vụ như một nền tảng cho các nghiên cứu khác liên quan đến an toàn giao thông, trí tuệ nhân tạo và học máy, giúp phát triển các mô hình dự đoán và cảnh báo sớm nhằm giảm thiểu số lượng tai nạn giao thông.

## **Phạm vi đề tài**

Đề tài "Xây dựng kho dữ liệu về tai nạn giao thông ở Mỹ" sẽ tập trung vào các nội dung và phạm vi chính sau đây:

**Phạm vi dữ liệu**: Kho dữ liệu sẽ thu thập thông tin từ các nguồn công khai, hợp pháp về tai nạn giao thông tại Mỹ, bao gồm:

* Thông tin về thời gian (năm, tháng, ngày, giờ) và địa điểm xảy ra tai nạn (bang, thành phố).
* Loại phương tiện và số lượng phương tiện tham gia tai nạn.
* Các đặc điểm liên quan đến người tham gia giao thông (độ tuổi, giới tính, tình trạng sức khỏe khi xảy ra tai nạn nếu có).

**Phạm vi công nghệ**: Đề tài sẽ áp dụng các công cụ và kỹ thuật kho dữ liệu phổ biến, bao gồm:

* Quy trình ETL (Extract, Transform, Load) để thu thập, làm sạch và chuyển đổi dữ liệu vào kho dữ liệu.
* Hệ quản trị cơ sở dữ liệu (DBMS) phù hợp để xây dựng và quản lý kho dữ liệu, tối ưu hóa truy vấn và đảm bảo khả năng mở rộng.
* Công cụ báo cáo và trực quan hóa dữ liệu, hỗ trợ người dùng cuối trong việc truy xuất và phân tích dữ liệu dễ dàng.

**Phạm vi phân tích**: Đề tài tập trung vào phân tích các xu hướng và mẫu hình tai nạn giao thông theo thời gian và không gian.

* Phân tích các yếu tố tác động đến tỷ lệ và mức độ nghiêm trọng của các vụ tai nạn.
* Cung cấp các bảng báo cáo, biểu đồ trực quan giúp dễ dàng theo dõi các chỉ số chính về tai nạn giao thông.

**Phạm vi ứng dụng**:

* Hệ thống sẽ phục vụ cho các đối tượng nghiên cứu và các cơ quan quản lý giao thông, nhằm hỗ trợ đưa ra các chính sách an toàn hiệu quả.
* Cung cấp một nền tảng dữ liệu để phát triển các mô hình dự đoán, nhằm cảnh báo và giảm thiểu nguy cơ tai nạn giao thông.

**Giới hạn của đề tài**:

* Dữ liệu chỉ bao gồm thông tin tại Mỹ và tập trung vào các vụ tai nạn đường bộ.
* Đề tài chỉ xây dựng kho dữ liệu ở mức độ cơ bản, chưa tích hợp các mô hình học máy hay dự báo phức tạp.
* Đề tài không đi sâu vào phân tích các yếu tố pháp lý hoặc các chính sách giao thông cụ thể.

## **Đối tượng sử dụng**

Kho dữ liệu về tai nạn giao thông tại Mỹ hướng đến phục vụ các đối tượng chính sau đây:

**Cơ quan quản lý giao thông và chính phủ**:

* + Các cơ quan quản lý giao thông như Bộ Giao thông Vận tải, cơ quan an toàn đường bộ, và chính quyền địa phương có thể sử dụng kho dữ liệu để phân tích các xu hướng tai nạn, từ đó xây dựng các chính sách, biện pháp và chiến dịch tuyên truyền nhằm giảm thiểu tai nạn giao thông.
  + Hỗ trợ trong việc cải thiện và điều chỉnh hạ tầng giao thông tại các khu vực có tỷ lệ tai nạn cao, góp phần nâng cao an toàn cho cộng đồng.

**Các nhà nghiên cứu và học giả**:

* + Các nhà nghiên cứu trong lĩnh vực giao thông, khoa học dữ liệu, và an toàn công cộng có thể sử dụng kho dữ liệu để phân tích sâu hơn về các yếu tố gây tai nạn và đề xuất các giải pháp giảm thiểu tai nạn giao thông.
  + Dữ liệu có thể làm nền tảng cho các nghiên cứu khoa học, mô hình dự báo và phân tích rủi ro, đóng góp vào kho tàng kiến thức về giao thông và an toàn đường bộ.

**Doanh nghiệp bảo hiểm**:

* + Các công ty bảo hiểm có thể khai thác kho dữ liệu để phân tích và đánh giá rủi ro tại các khu vực và nhóm đối tượng cụ thể, từ đó xây dựng chính sách bảo hiểm phù hợp, quản lý rủi ro và đưa ra các khuyến nghị cho khách hàng.

**Các tổ chức phi chính phủ (NGO) về an toàn giao thông**:

* + Các tổ chức phi lợi nhuận và các tổ chức phi chính phủ hoạt động trong lĩnh vực an toàn giao thông có thể sử dụng dữ liệu để thiết kế các chương trình và hoạt động nhằm nâng cao ý thức tham gia giao thông, bảo vệ người tham gia giao thông và giảm thiểu thiệt hại do tai nạn gây ra.

**Những người làm công nghệ và chuyên gia phân tích dữ liệu**:

* + Các chuyên gia và kỹ sư dữ liệu có thể sử dụng kho dữ liệu làm nguồn tài nguyên cho việc xây dựng, thử nghiệm các công cụ phân tích và mô hình dữ liệu lớn, từ đó ứng dụng trong các dự án về phân tích và dự đoán giao thông, góp phần phát triển công nghệ trong lĩnh vực quản lý dữ liệu giao thông.

## **Phương pháp nghiên cứu**

Để xây dựng kho dữ liệu về tai nạn giao thông ở Mỹ một cách khoa học và hiệu quả, đề tài áp dụng các phương pháp nghiên cứu sau:

1. **Phương pháp thu thập dữ liệu**:
   * **Thu thập từ các nguồn công khai**: Dữ liệu sẽ được thu thập từ các nguồn dữ liệu chính thống và công khai, như các cơ quan giao thông Mỹ (ví dụ: NHTSA - Cơ quan An toàn Giao thông Đường bộ Quốc gia) và các trang web cung cấp dữ liệu mở (open data).
   * **Chuẩn hóa dữ liệu**: Trong quá trình thu thập, dữ liệu từ nhiều nguồn sẽ được chuẩn hóa nhằm đảm bảo tính nhất quán và thống nhất về cấu trúc, giúp dễ dàng trong quá trình tích hợp vào kho dữ liệu.
2. **Phương pháp xử lý dữ liệu (ETL - Extract, Transform, Load)**:
   * **Trích xuất dữ liệu (Extract)**: Dữ liệu được thu thập từ các tệp CSV, cơ sở dữ liệu công khai hoặc các API cung cấp dữ liệu về tai nạn giao thông.
   * **Chuyển đổi dữ liệu (Transform)**: Dữ liệu sau khi được trích xuất sẽ được làm sạch, chuẩn hóa và chuyển đổi để đảm bảo tính toàn vẹn và loại bỏ các giá trị thiếu, lỗi hoặc không nhất quán. Ngoài ra, các thuộc tính bổ sung như ngày, giờ, và địa điểm có thể được chuyển đổi để phù hợp với các phân tích trong kho dữ liệu.
   * **Tải dữ liệu (Load)**: Dữ liệu đã qua xử lý sẽ được tải vào kho dữ liệu theo một cấu trúc hợp lý, đáp ứng nhu cầu phân tích và truy vấn sau này.
3. **Phương pháp thiết kế và xây dựng kho dữ liệu**:
   * **Thiết kế mô hình dữ liệu**: Sử dụng các phương pháp thiết kế mô hình dữ liệu, tập trung vào mô hình ngôi sao hoặc mô hình tuyết (star schema hoặc snowflake schema), để đảm bảo dữ liệu được tổ chức một cách khoa học, dễ dàng truy vấn và phân tích.
   * **Tối ưu hóa kho dữ liệu**: Tối ưu hóa cấu trúc kho dữ liệu để cải thiện hiệu suất truy vấn và đáp ứng nhu cầu mở rộng trong tương lai.
4. **Phương pháp phân tích dữ liệu**:
   * **Phân tích mô tả (Descriptive Analysis)**: Thực hiện các phân tích mô tả để xác định xu hướng, mẫu hình và các chỉ số quan trọng liên quan đến tai nạn giao thông như tần suất tai nạn, loại phương tiện thường gặp, và các yếu tố thời gian và địa điểm.
   * **Trực quan hóa dữ liệu**: Sử dụng các công cụ trực quan hóa (như Power BI, Tableau hoặc Python) để biểu diễn các mẫu hình và xu hướng tai nạn giao thông dưới dạng biểu đồ, bảng biểu, giúp người dùng dễ dàng hiểu và đưa ra nhận định.
5. **Phương pháp kiểm thử và đánh giá**:
   * **Kiểm thử hệ thống**: Kiểm thử các chức năng của kho dữ liệu, bao gồm khả năng truy xuất dữ liệu, tốc độ phản hồi và độ chính xác của các kết quả truy vấn, nhằm đảm bảo kho dữ liệu hoạt động ổn định.
   * **Đánh giá hiệu quả**: Dựa trên các yêu cầu ban đầu, tiến hành đánh giá mức độ đáp ứng của kho dữ liệu, khả năng hỗ trợ phân tích và ứng dụng vào thực tế, đồng thời ghi nhận các phản hồi từ người dùng cuối để cải thiện hệ thống.

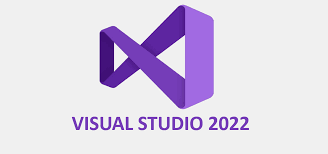
## **Bố cục báo cáo**

* Chương 1 Tổng quan
* Chương 2 Cơ sở lý thuyết
* Chương 3 Phân tích và thiết kế kho dữ liệu
* Chương 4 Quá trình ETL
* Kết Luận

# **CƠ SỞ LÝ THUYẾT**

## **Visual Studio**

**Visual Studio** là một môi trường phát triển tích hợp (IDE) do Microsoft phát triển, được sử dụng để phát triển nhiều loại ứng dụng, từ ứng dụng Windows truyền thống, ứng dụng di động, đến ứng dụng web và dịch vụ đám mây. Đây là công cụ hỗ trợ các ngôn ngữ lập trình như C++, C#, Visual Basic, Python, JavaScript, TypeScript, và nhiều ngôn ngữ khác thông qua các tiện ích mở rộng.Một socket được coi như một đầu cuối của một kết nối mạng, và nó chịu trách nhiệm cho việc gửi và nhận dữ liệu. Một socket được xác định bởi một địa chỉ IP và một số cổng (port). Địa chỉ IP xác định máy tính hoặc thiết bị mạng trong mạng lưới, và số cổng xác định ứng dụng đang chạy trên máy tính đó.



Visual Studio là một IDE cung cấp một nền tảng duy nhất cho các nhà phát triển để tạo ra nhiều loại ứng dụng khác nhau, bao gồm:

* **Ứng dụng Windows**: Phát triển các ứng dụng desktop, ứng dụng Universal Windows Platform (UWP), hoặc Windows Forms.
* **Ứng dụng Web**: Xây dựng các ứng dụng web sử dụng công nghệ như ASP.NET, HTML, CSS, JavaScript và nhiều framework khác như Angular, React, hoặc Vue.
* **Ứng dụng di động**: Visual Studio hỗ trợ phát triển ứng dụng di động cho cả Android và iOS thông qua Xamarin hoặc React Native, giúp lập trình viên có thể xây dựng các ứng dụng đa nền tảng.
* **Ứng dụng đám mây**: Visual Studio cung cấp các công cụ và tích hợp chặt chẽ với Microsoft Azure, giúp phát triển và triển khai các ứng dụng đám mây, API, và dịch vụ microservices.
* **Ứng dụng game**: Hỗ trợ phát triển game với Unity và Unreal Engine, bao gồm cả khả năng biên dịch và gỡ lỗi.
* **Ứng dụng AI và Machine Learning**: Tích hợp với các thư viện và dịch vụ AI, giúp tạo và triển khai các mô hình học máy và trí tuệ nhân tạo.
* **Ứng dụng IoT (Internet of Things)**: Hỗ trợ phát triển các ứng dụng IoT, từ các thiết bị nhúng nhỏ đến các hệ thống IoT lớn tích hợp đám mây.

**Các tính năng chính của Visual Studio:**

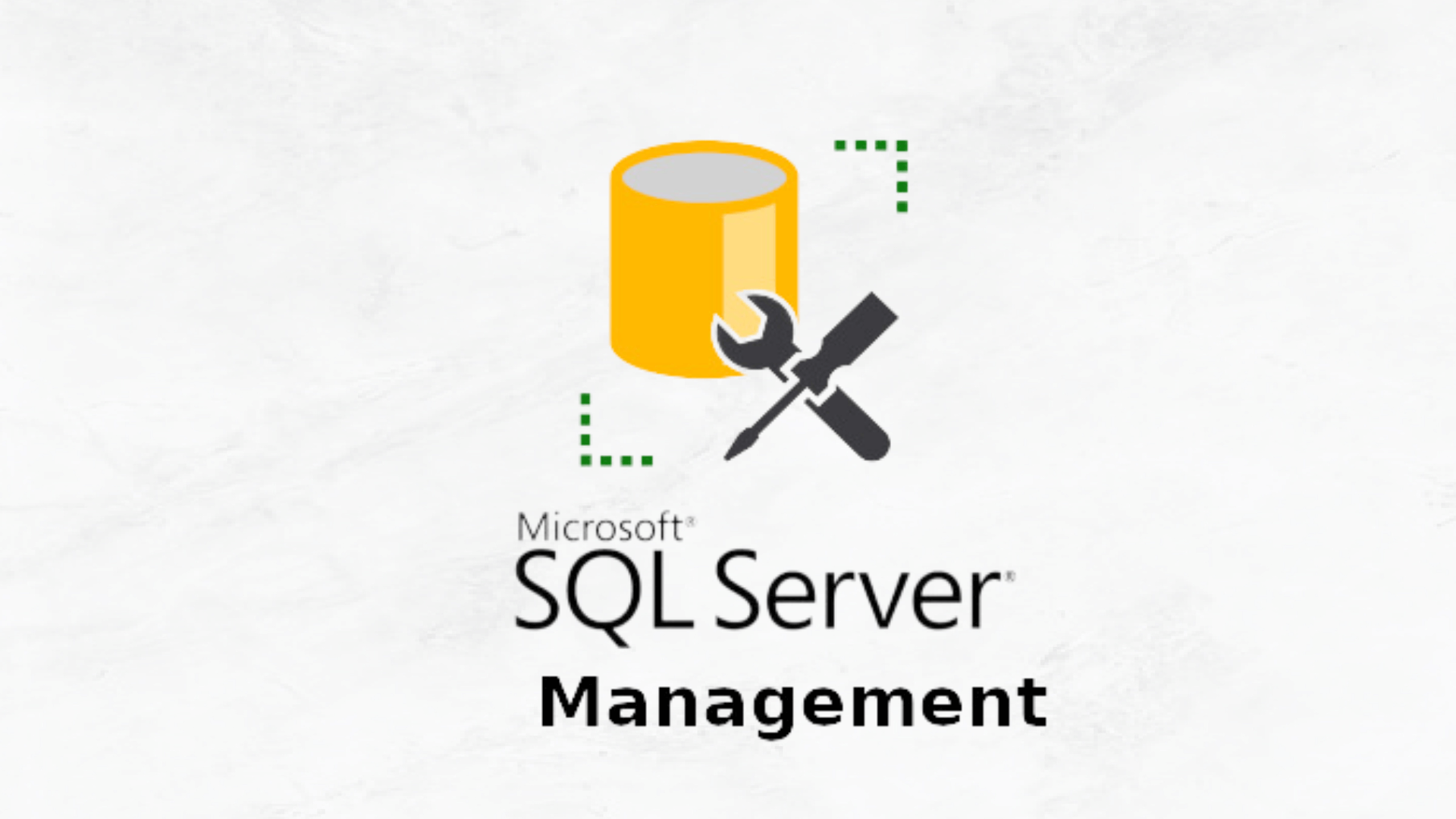
* **Trình gỡ lỗi (Debugger)**: Tích hợp sẵn để kiểm tra và sửa lỗi trong mã nguồn.
* **Trình chỉnh sửa mã nguồn thông minh**: Có khả năng gợi ý mã, tự động hoàn thành mã (IntelliSense), và kiểm tra cú pháp.
* **Tích hợp Git**: Giúp người dùng dễ dàng quản lý mã nguồn, làm việc với các nhánh và đồng bộ hóa với các kho lưu trữ từ xa như GitHub và Azure DevOps.
* **Hỗ trợ phát triển ứng dụng đa nền tảng**: Visual Studio hỗ trợ phát triển ứng dụng cho nhiều nền tảng khác nhau như Android, iOS, macOS, và Windows.
* **Tích hợp với các dịch vụ Azure**: Giúp các lập trình viên triển khai và quản lý ứng dụng trực tiếp trên nền tảng đám mây Azure của Microsoft.

Sử dụng **Visual Studio** mang lại rất nhiều lợi ích cho các nhà phát triển, đặc biệt là những ai làm việc với các dự án phức tạp và đa nền tảng. Dưới đây là một số lợi ích chính:

* **Hỗ trợ nhiều ngôn ngữ lập trình**: Visual Studio hỗ trợ nhiều ngôn ngữ như C#, C++, Python, JavaScript, TypeScript, và nhiều ngôn ngữ khác. Điều này giúp nhà phát triển linh hoạt trong việc chọn ngôn ngữ phù hợp với dự án của mình.
* **Trình gợi ý mã thông minh (IntelliSense)**: Tính năng IntelliSense tự động hoàn thành mã và gợi ý khi nhập, giúp tăng tốc độ viết mã, giảm lỗi và giúp nhà phát triển dễ dàng hiểu cấu trúc và hàm của mã.
* **Trình gỡ lỗi mạnh mẽ**: Trình gỡ lỗi tích hợp giúp phát hiện và sửa lỗi dễ dàng, hỗ trợ cả gỡ lỗi trực tiếp và từ xa, giúp tiết kiệm thời gian và giảm bớt khó khăn trong quá trình phát triển.
* **Tích hợp Git và quản lý mã nguồn**: Visual Studio có tích hợp sẵn Git giúp quản lý mã nguồn và làm việc với các kho lưu trữ như GitHub, Azure DevOps. Tính năng này rất tiện lợi cho làm việc nhóm và theo dõi lịch sử mã nguồn.
* **Phát triển đa nền tảng**: Với Visual Studio, nhà phát triển có thể tạo ra ứng dụng chạy trên nhiều hệ điều hành như Windows, macOS, Android, và iOS từ một mã nguồn duy nhất. Điều này rất hữu ích khi phát triển các ứng dụng đa nền tảng.
* **Hỗ trợ dịch vụ đám mây và DevOps**: Tích hợp với Azure giúp triển khai và quản lý các ứng dụng đám mây. Ngoài ra, Visual Studio còn hỗ trợ các quy trình DevOps như CI/CD (Continuous Integration/Continuous Delivery), giúp triển khai ứng dụng một cách tự động và nhanh chóng.
* **Nhiều công cụ kiểm thử**: Visual Studio cung cấp các công cụ kiểm thử như unit test, kiểm thử UI, và các bộ công cụ kiểm tra hiệu suất, giúp nâng cao chất lượng của ứng dụng trước khi phát hành.
* **Kho tiện ích mở rộng phong phú**: Visual Studio có kho tiện ích mở rộng đa dạng, giúp nhà phát triển tùy chỉnh IDE để phù hợp với nhu cầu cụ thể của dự án, từ hỗ trợ ngôn ngữ đến công cụ năng suất.
* **Cộng đồng lớn và tài liệu phong phú**: Visual Studio có một cộng đồng phát triển lớn và nhiều tài liệu hỗ trợ, vì vậy nhà phát triển dễ dàng tìm kiếm tài liệu và hỗ trợ khi cần.
* **Phiên bản miễn phí và dễ tiếp cận**: Phiên bản Visual Studio Community là miễn phí, giúp các cá nhân và nhóm nhỏ có thể sử dụng mà không phải lo ngại về chi phí.

## **SSMS**

**SSMS** (SQL Server Management Studio) là một công cụ quản lý cơ sở dữ liệu do Microsoft phát triển, dùng để làm việc với SQL Server – một hệ quản trị cơ sở dữ liệu phổ biến. SSMS là môi trường tích hợp (IDE) cho phép các quản trị viên cơ sở dữ liệu và nhà phát triển dễ dàng truy cập, quản lý, cấu hình, và phát triển cơ sở dữ liệu SQL Server.



**Các tính năng nổi bật của SSMS:**

* Quản lý cơ sở dữ liệu SQL Server: SSMS cung cấp giao diện người dùng trực quan để quản lý các phiên bản SQL Server. Người dùng có thể thực hiện các thao tác như sao lưu, khôi phục, sao chép cơ sở dữ liệu, và cấu hình bảo mật.
* Trình chỉnh sửa truy vấn SQL: SSMS có trình chỉnh sửa mã hỗ trợ viết và chạy các truy vấn SQL, giúp người dùng dễ dàng thao tác với dữ liệu. Ngoài ra, trình chỉnh sửa còn có tính năng kiểm tra cú pháp, gợi ý mã và phân tích hiệu suất truy vấn.
* Tạo và quản lý các đối tượng cơ sở dữ liệu: Người dùng có thể tạo và quản lý các bảng, stored procedures, views, triggers, và các đối tượng khác trong cơ sở dữ liệu.
* Quản lý bảo mật: SSMS cho phép quản lý quyền truy cập và vai trò người dùng, từ đó thiết lập các chính sách bảo mật và kiểm soát truy cập cơ sở dữ liệu.
* Tích hợp công cụ báo cáo: SSMS hỗ trợ các công cụ báo cáo và tính năng báo cáo hiệu suất, giúp các quản trị viên theo dõi, phân tích và tối ưu hóa hệ thống SQL Server.
* Tính năng gỡ lỗi: SSMS cung cấp tính năng gỡ lỗi để theo dõi các lỗi trong các stored procedure và các đoạn mã T-SQL phức tạp.
* Tích hợp với Azure: SSMS hỗ trợ kết nối với các dịch vụ đám mây của Microsoft Azure, như Azure SQL Database, giúp quản lý cơ sở dữ liệu đám mây dễ dàng từ cùng một giao diện.

**Lợi ích khi sử dụng SSMS:**

* Tăng năng suất: Cung cấp giao diện quản lý trực quan, giúp thao tác nhanh chóng và dễ dàng.
* Quản lý toàn diện: SSMS có các công cụ và tiện ích giúp quản lý toàn diện SQL Server, từ cấu hình đến bảo mật.
* Hỗ trợ gỡ lỗi và tối ưu hóa: Các công cụ tích hợp giúp dễ dàng phát hiện lỗi và tối ưu hóa hiệu suất cơ sở dữ liệu.

## **SSIS**

**SSIS** (SQL Server Integration Services) là một công cụ tích hợp dữ liệu và thực thi các quy trình ETL (Extract, Transform, Load - Trích xuất, Chuyển đổi, Tải) do Microsoft phát triển, được sử dụng để xử lý, di chuyển và biến đổi dữ liệu trong SQL Server. SSIS là một phần của SQL Server và đóng vai trò quan trọng trong việc xây dựng và quản lý các quy trình dữ liệu phức tạp, thường được sử dụng trong các hệ thống kho dữ liệu (data warehouse) và hệ thống phân tích.



**Các tính năng chính của SSIS:**

* ETL (Extract, Transform, Load): SSIS hỗ trợ các quy trình ETL giúp trích xuất dữ liệu từ nhiều nguồn (như cơ sở dữ liệu, tệp văn bản, API), chuyển đổi (xử lý) dữ liệu theo các quy tắc kinh doanh, và tải dữ liệu vào kho dữ liệu hoặc các hệ thống lưu trữ khác.
* Tích hợp dữ liệu từ nhiều nguồn: SSIS có khả năng kết nối với nhiều loại nguồn dữ liệu khác nhau, bao gồm SQL Server, Oracle, MySQL, các tệp CSV, Excel, và nhiều nguồn dữ liệu đám mây, giúp tích hợp và hợp nhất dữ liệu từ nhiều nguồn khác nhau.
* Làm sạch và chuyển đổi dữ liệu: SSIS cung cấp các công cụ giúp làm sạch dữ liệu, xử lý dữ liệu thiếu, chuẩn hóa định dạng và thực hiện các quy trình chuyển đổi phức tạp trước khi đưa vào kho dữ liệu.
* Tự động hóa quy trình dữ liệu: Với SSIS, người dùng có thể tự động hóa các quy trình dữ liệu, từ chuyển đổi dữ liệu đến tải dữ liệu vào kho, qua đó tiết kiệm thời gian và giảm thiểu lỗi thủ công.
* Kiểm soát luồng công việc: SSIS cung cấp các điều kiện và điều khiển luồng công việc, cho phép xây dựng các quy trình phức tạp với các điều kiện thực hiện, vòng lặp và xử lý song song.
* Xử lý sự kiện và lỗi: SSIS hỗ trợ xử lý lỗi và sự kiện, giúp phát hiện và khắc phục sự cố trong quá trình tải dữ liệu.
* Hỗ trợ lập lịch và giám sát: SSIS có thể tích hợp với SQL Server Agent để lập lịch các gói công việc, cho phép quản lý và giám sát các quy trình ETL một cách tự động.

**Lợi ích của SSIS:**

* Hiệu quả và tự động hóa quy trình dữ liệu: SSIS giúp tự động hóa các tác vụ dữ liệu, từ đó giảm thời gian và công sức xử lý.
* Xử lý dữ liệu lớn: SSIS được thiết kế để xử lý lượng dữ liệu lớn với hiệu suất cao, đặc biệt hữu ích trong các kho dữ liệu và hệ thống phân tích.
* Khả năng tích hợp rộng: SSIS có thể làm việc với nhiều loại nguồn và đích dữ liệu, hỗ trợ khả năng tích hợp cao.
* Giảm thiểu lỗi: Bằng cách tự động hóa và kiểm soát luồng công việc, SSIS giúp giảm thiểu các lỗi thủ công và đảm bảo tính nhất quán của dữ liệu.

## **SSAS**

**SSAS** (SQL Server Analysis Services) là một công cụ phân tích dữ liệu và xử lý OLAP (Online Analytical Processing) do Microsoft phát triển, được tích hợp trong SQL Server. SSAS đóng vai trò quan trọng trong việc xây dựng các hệ thống phân tích và báo cáo dữ liệu mạnh mẽ, thường được sử dụng trong các hệ thống kho dữ liệu và các giải pháp Business Intelligence (BI).



**Các Tính Năng Chính của SSAS**:

* SSAS hỗ trợ tạo các mô hình OLAP, cho phép người dùng thực hiện phân tích đa chiều và khám phá dữ liệu ở các mức độ chi tiết khác nhau.
* Các mô hình OLAP trong SSAS được tổ chức dưới dạng các cube, cho phép xử lý và phân tích dữ liệu từ nhiều góc độ (chiều) như thời gian, địa lý, sản phẩm, v.v.
* SSAS cung cấp các thuật toán và công cụ khai phá dữ liệu, giúp phát hiện các mẫu hình, xu hướng và mối quan hệ trong dữ liệu.
* Từ đó, người dùng có thể áp dụng các kỹ thuật khai phá dữ liệu để dự đoán và đưa ra các quyết định kinh doanh chiến lược.
* Bên cạnh mô hình OLAP truyền thống, SSAS còn hỗ trợ mô hình bảng (Tabular Model) dựa trên bộ nhớ, cho phép tạo các mô hình phân tích dữ liệu nhanh và hiệu quả.
* Mô hình bảng thích hợp cho các hệ thống phân tích lớn với khả năng tính toán nhanh chóng nhờ vào công nghệ lưu trữ dữ liệu trong bộ nhớ RAM.
* SSAS cho phép quản lý bảo mật và phân quyền truy cập, giúp kiểm soát quyền hạn của người dùng đối với từng chiều hoặc thành phần của dữ liệu.
* Điều này đảm bảo tính bảo mật và giới hạn truy cập vào các dữ liệu nhạy cảm.
* SSAS dễ dàng tích hợp với các công cụ BI khác của Microsoft như Power BI, Excel, và các công cụ BI bên thứ ba khác.
* Khả năng tích hợp này giúp dễ dàng xây dựng báo cáo và bảng điều khiển trực quan, từ đó hỗ trợ người dùng cuối trong việc phân tích và ra quyết định.
* SSAS được thiết kế để tối ưu hóa hiệu suất truy vấn, hỗ trợ xử lý dữ liệu lớn một cách hiệu quả.
* Khả năng mở rộng cao giúp SSAS đáp ứng nhu cầu của các hệ thống BI lớn và phức tạp.

**Lợi Ích của SSAS**:

* Với các mô hình phân tích đa chiều và bảng, SSAS cung cấp khả năng phân tích nhanh chóng và chính xác, giúp người dùng hiểu rõ hơn về các mẫu hình và xu hướng trong dữ liệu.
* SSAS giúp tạo ra các báo cáo và bảng điều khiển dựa trên dữ liệu thực tế, cung cấp thông tin cần thiết cho việc đưa ra quyết định kinh doanh sáng suốt.
* Khả năng tương tác và tích hợp với các công cụ BI như Power BI, Excel giúp người dùng cuối dễ dàng truy cập và làm việc với dữ liệu phân tích mà không cần quá nhiều kiến thức kỹ thuật.
* Với các tính năng quản lý bảo mật và phân quyền, SSAS đảm bảo rằng dữ liệu được bảo vệ và chỉ có người dùng được phép mới có thể truy cập vào dữ liệu quan trọng.
* SSAS cung cấp các công cụ khai phá dữ liệu tiên tiến, giúp người dùng không chỉ phân tích dữ liệu mà còn khám phá các thông tin ẩn sâu trong dữ liệu, hỗ trợ việc dự đoán và tối ưu hóa kinh doanh.

## **SSRS**

**SSRS** (SQL Server Reporting Services) là một công cụ tạo báo cáo mạnh mẽ do Microsoft phát triển, dùng để thiết kế, phát triển và phân phối các báo cáo dựa trên dữ liệu trong SQL Server. SSRS là một phần của bộ SQL Server, giúp các doanh nghiệp có thể tạo ra các báo cáo chi tiết, trực quan và dễ dàng chia sẻ với người dùng.



**Các tính năng chính của SSRS:**

* Tạo và thiết kế báo cáo: SSRS cho phép tạo các báo cáo đa dạng như báo cáo bảng, báo cáo đồ thị, báo cáo ma trận, và báo cáo tự do, giúp đáp ứng nhu cầu phân tích thông tin đa dạng.
* Truy cập dữ liệu từ nhiều nguồn: SSRS không chỉ làm việc với SQL Server mà còn có thể kết nối với nhiều loại nguồn dữ liệu khác như Oracle, Azure SQL Database, Excel, và nhiều loại cơ sở dữ liệu khác.
* Tùy chỉnh và tương tác với báo cáo: SSRS hỗ trợ tùy chỉnh báo cáo theo yêu cầu và cho phép người dùng tương tác với báo cáo bằng cách lọc, sắp xếp, và xuất dữ liệu.
* Lập lịch và phân phối báo cáo: SSRS có thể tích hợp với SQL Server Agent để lập lịch phát hành báo cáo tự động và phân phối qua email, trang web, hoặc lưu trữ trên máy chủ.
* Khả năng xuất báo cáo đa dạng: Báo cáo từ SSRS có thể được xuất ra nhiều định dạng như PDF, Excel, Word, HTML, và các định dạng tệp khác, giúp dễ dàng chia sẻ và tích hợp với các hệ thống khác.
* Quản lý và bảo mật: SSRS cung cấp công cụ quản lý người dùng và quyền truy cập, giúp bảo mật dữ liệu và chỉ cho phép người dùng có quyền truy cập vào các báo cáo cụ thể.
* Tích hợp với các công cụ Microsoft khác: SSRS có thể tích hợp với Power BI, SharePoint và các dịch vụ Azure, giúp tăng cường khả năng phân tích và báo cáo trong môi trường Microsoft.

**Lợi ích của SSRS:**

* Trực quan hóa dữ liệu: SSRS cung cấp nhiều loại báo cáo và biểu đồ, giúp người dùng dễ dàng hiểu và phân tích dữ liệu.
* Tăng năng suất: Tự động hóa việc tạo và phân phối báo cáo giúp tiết kiệm thời gian và tăng năng suất cho bộ phận phân tích và quản lý.
* Khả năng tùy chỉnh cao: SSRS cho phép tạo các báo cáo linh hoạt, tùy chỉnh theo nhu cầu riêng của từng người dùng hoặc phòng ban.
* Tiết kiệm chi phí: SSRS là một công cụ báo cáo tích hợp trong SQL Server, không yêu cầu các phần mềm báo cáo bên ngoài, giúp tiết kiệm chi phí cho doanh nghiệp.

# **PHÂN TÍCH THIẾT KẾ KHO DỮ LIỆU**

## **Mô hình logic**

### ***Dimensions:***

**Dim\_Date**

| Dim\_Date |
| --- |
| DateID  DateTime  Year  Month  Day  Hour  Minute |

**Dim\_Location**

| Dim\_Location |
| --- |
| LocationID  Street  City  County  State  Zipcode  Country |

**Dim\_RoadFeature**

| Dim\_RoadFeature |
| --- |
| RoadFeatureID  Amenity  Bump  Give\_Away  Crossing  Junction  No\_Exit  Railway  Roundabout  Station  Stop  Traffic\_Signal  Traffic\_Calming  Turning\_Loop |

**Dim\_Twilight**

| Dim\_Twilight |
| --- |
| TwilightID  Nautical\_Twilight  Civil\_Twilight  Sunrise\_Sunset  Astronomical\_Twilight |

**Dim\_Weather**

| Dim\_Weather |
| --- |
| WeatherID  Temperature(F)  Wind\_Chill(F)  Humidity(%)  Pressure(in)  Visibility(mi)  Wind\_Direction  Win\_Speed(mph)  Precipitation(in)  Weather\_Condition |

**Dim\_Driver**

| Dim\_Driver |
| --- |
| DriverID  Age\_Brand\_of\_Driver  Age\_of\_Vehicle  Driver\_Home\_Area\_Type  Driver\_IMD\_Delice  Sex\_of\_Driver |

**Dim\_TypeImpact**

| Dim\_TypeImpact |
| --- |
| TypeImpactID  X1st\_Point\_of\_Impact  Vehicle\_Manoeuvre |

**Dim\_Vehicle**

| Dim\_Vehicle |
| --- |
| VehicleID  make  model  Vehicle\_Type  Propulsion\_Code |

### ***Facts:***

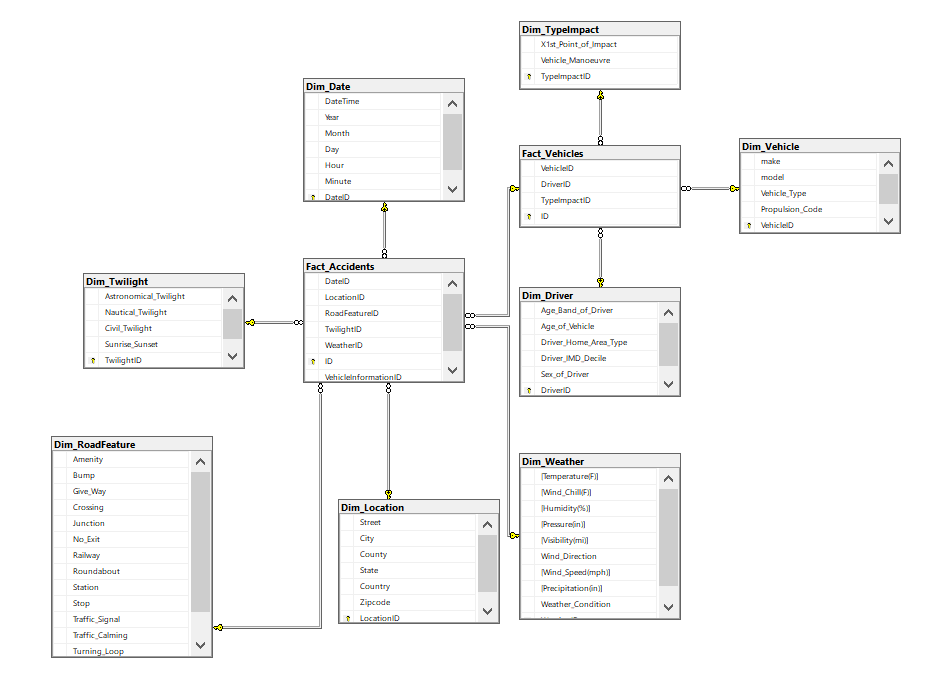
**Fact\_Vehicles**

| Fact\_Vehicles |
| --- |
| ID  VehicleID  DriverID  TypeImpactID |

**Fact\_Accidents**

| Fact\_Accidents |
| --- |
| ID  VehicleInformationID  DateID  LocationID  WeatherID  TwilightID  RoadFeatureID |

### ***Sơ đồ mô hình logic:***



## **Mô hình hóa vật lý**

### ***Dimensions:***

**Dim\_Date**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| DateID | int | Chỉ mục khóa chính |
| DateTime | nvarchar(255) | Thời gian xảy ra tai nạn (YYYY/MM/DD HH:MM:SS) |
| Year | nvarchar(255) | Năm xảy ra tai nạn |
| Month | nvarchar(255) | Tháng xảy ra tai nạn (1-12) |
| Day | nvarchar(255) | Ngày xảy ra tai nạn (1-31) |
| Hour | nvarchar(255) | Giờ xảy ra tai nạn (0-23) |
| Minute | nvarchar(255) | Phút xảy ra tai nạn (0-59) |

**Dim\_Location**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| LocationID | INT | Chỉ mục khóa chính địa điểm |
| Street | nvarchar(255) | Tên đường |
| City | nvarchar(255) | Tên thành phố |
| County | nvarchar(255) | Tên quận |
| State | nvarchar(255) | Tên tiểu bang |
| Contry | nvarchar(255) | Tên quốc gia |
| Zipcode | nvarchar(255) | Mã khu vực |

**Dim\_RoadFeature**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| RoadFeatureID | INT | Chỉ mục khóa chỉnh đặc điểm đường |
| Amenity | Boolean | Tiện nghi |
| Bump | Boolean | Sốc |
| Give\_Away | Boolean | Cho đi |
| Crossing | Boolean | Qua đường |
| Junction | Boolean | Ngã ba |
| No\_Exit | Boolean | Lối thoát |
| Railway | Boolean | Đường sắt |
| Roundabout | Boolean | Bùng binh |
| Station | Boolean | Ga tàu |
| Stop | Boolean | Dừng lại |
| Traffic\_Signal | Boolean | Tín hiệu giao thông |
| Traffic\_Calming | Boolean | Làm dịu giao thông |
| Turning\_Loop | Boolean | Vòng xoay |

**Dim\_Twilight**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| TwilightID | INT | Chỉ mục khóa chính hoàng hôm |
| Astronomical | nvarchar(255) | Thiên văn |
| Nautical | nvarchar(255) | Hàng hải |
| Civil | nvarchar(255) | Chạng vạng |
| Sunrise | nvarchar(255) | Bình minh |

**Dim\_Weather**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| WeatherID | INT | Chỉ mục khóa chính thời tiết |
| Temperature | float | Nhiệt độ |
| Wind\_Chill | float | Gió lạnh |
| Humidity | float | Độ ẩm |
| Pressure | float | Áp suất |
| Visibility | float | Tầm nhìn |
| Wind\_Direction | nvarchar(255) | Hướng gió |
| Wind\_Speed | float | Tốc độ gió |
| Precipitation | float | Lượng mưa |
| Weather\_Condition | nvarchar(255) | Điều kiện thời tiết |

**Dim\_Vehicle**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| VehicleID | INT | Chỉ mục khóa chính phương tiện |
| make | nvarchar(255) | Hãng xe |
| model | nvarchar(255) | Mẫu xe |
| Vehicle\_Type | nvarchar(255) | Kiểu xe |
| Propulsion\_Code | nvarchar(255) | Mã động lực |

**Dim\_Driver**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| DriverID | INT | Chỉ mục khóa chính người điều khiển |
| Age\_Brand\_of\_Driver | nvarchar(255) | Nhóm tuổi tài xế |
| Age\_of\_Vehicle | nvarchar(255) | Tuổi của xe |
| Sex\_of\_Driver | nvarchar(255) | Giới tính tài xế |
| Driver\_Home | nvarchar(255) | Nơi ở của tài xế |
| Driver\_IMD\_Delice | float | Chỉ số rối loạn bẩm sinh (1-10) |

**Dim\_TypeImpact**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| TypeImpactID | INT | Chỉ mục khóa chính kiểu va chạm |
| X1st\_Point\_of\_Impact | nvarchar(255) | Hướng, điểm va chạm (right, left,...) |
| Vehicle\_Manouvre | nvarchar(255) | Trạng thái điều khiển xe |

### ***Facts:***

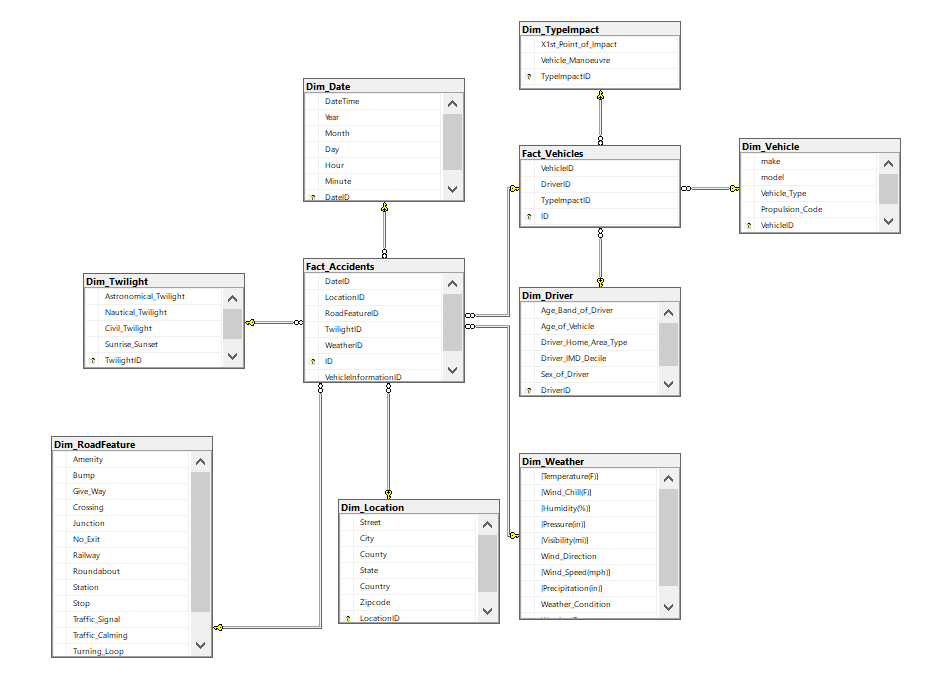
**Fact\_Accidents**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| ID | INT | Chỉ mục khóa chính vụ tai nạn |
| DateID | INT | Chỉ mục khóa ngoại thời gian tai nạn |
| LocationID | INT | Chỉ mục khóa ngoại địa điểm tai nạn |
| WeatherID | INT | Chỉ mục khóa ngoại thời tiết |
| TwilightID | INT | Chỉ mục khóa ngoại chạng vạng |
| RoadFeatureID | INT | Chỉ mục khóa ngoại đặc điểm đường |
| VehicleInformationID | INT | Chỉ mục khóa ngoại đặc điểm phương tiện tai nạn |

**Fact\_Vehicles**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| ID | INT | Chỉ mục khóa chính phương tiện |
| VehicleID | INT | Chỉ mục khóa ngoại phương tiện |
| TypeImpactID | INT | Chỉ mục khóa ngoại kiểu tai nạn |
| DriverID | INT | Chỉ mục khóa ngoại tài xế |

### ***Sơ đồ mô hình hóa vật lý:***

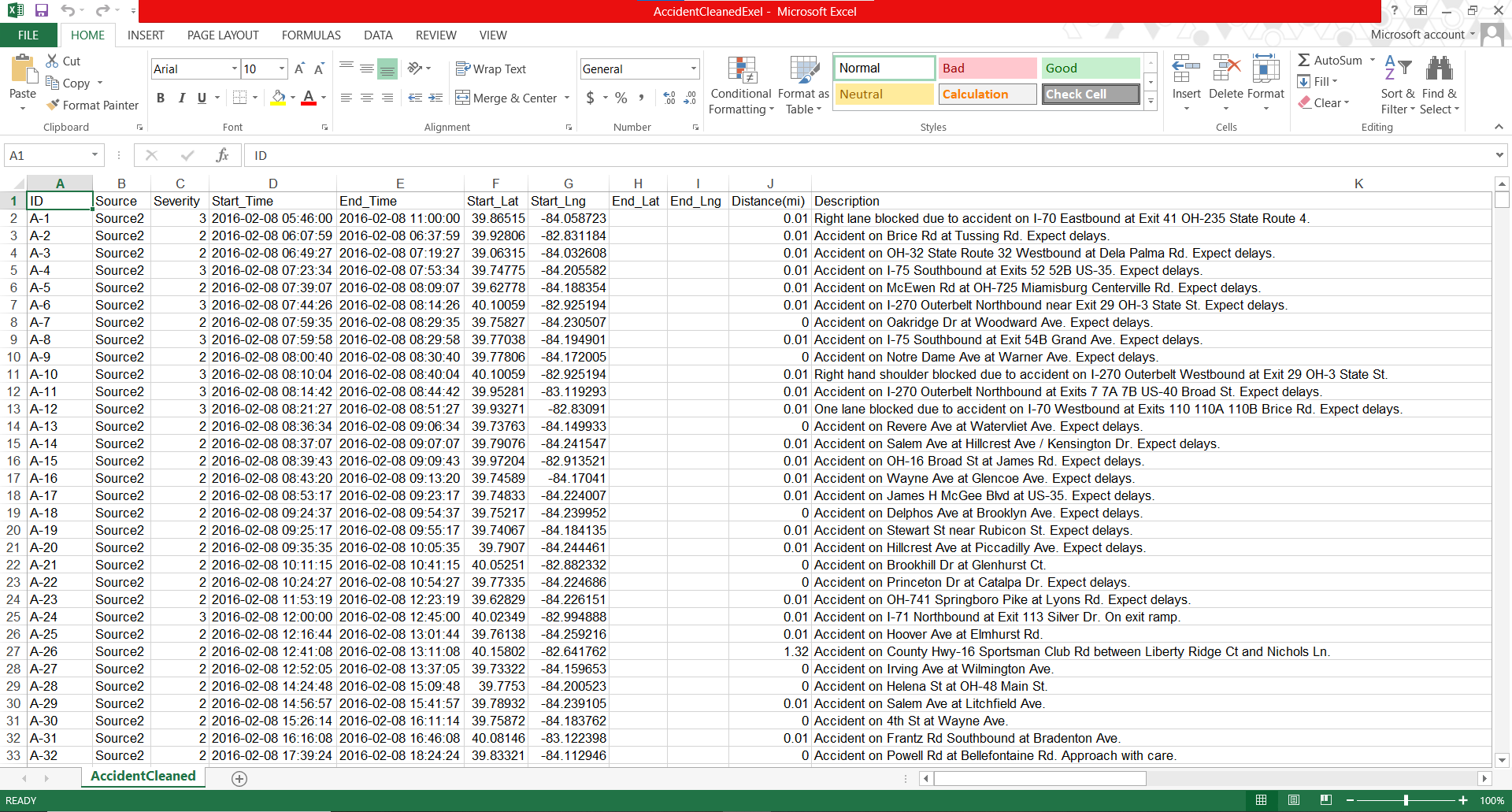


# **QUÁ TRÌNH ETL**

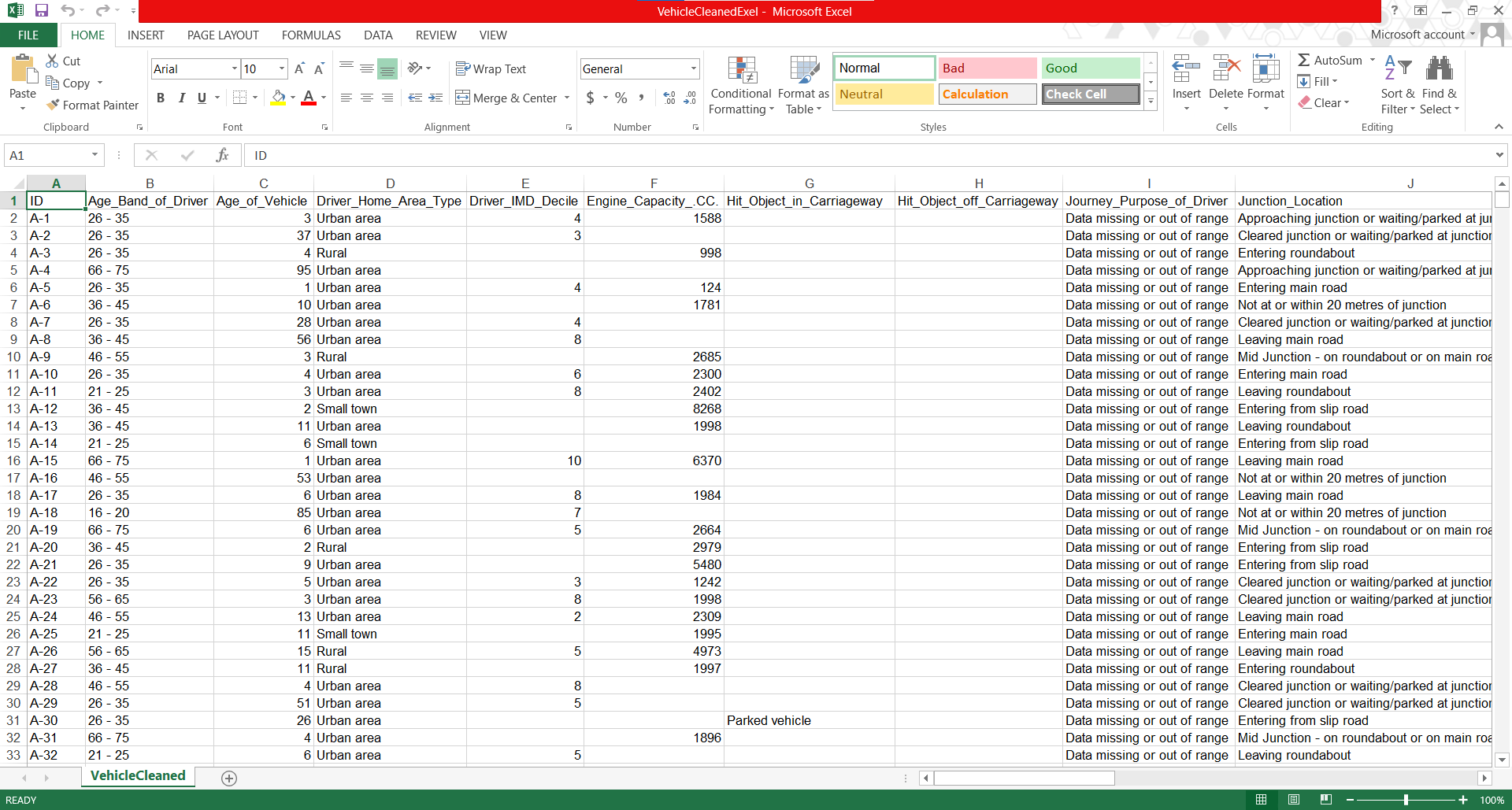
## **Chuẩn bị dữ liệu**

Nguồn dữ liệu bao gồm 2 file exel hoặc csv:

File accidents

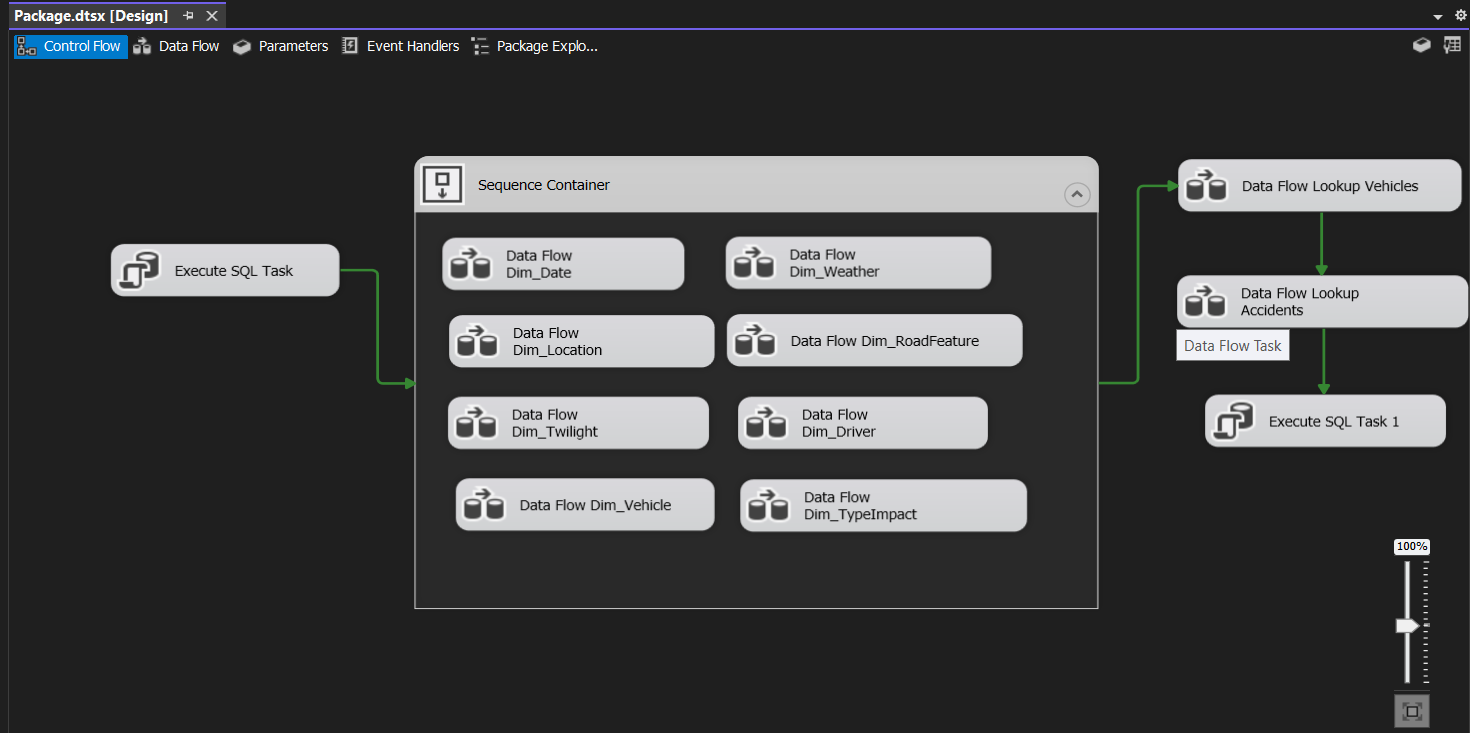


File vehicles



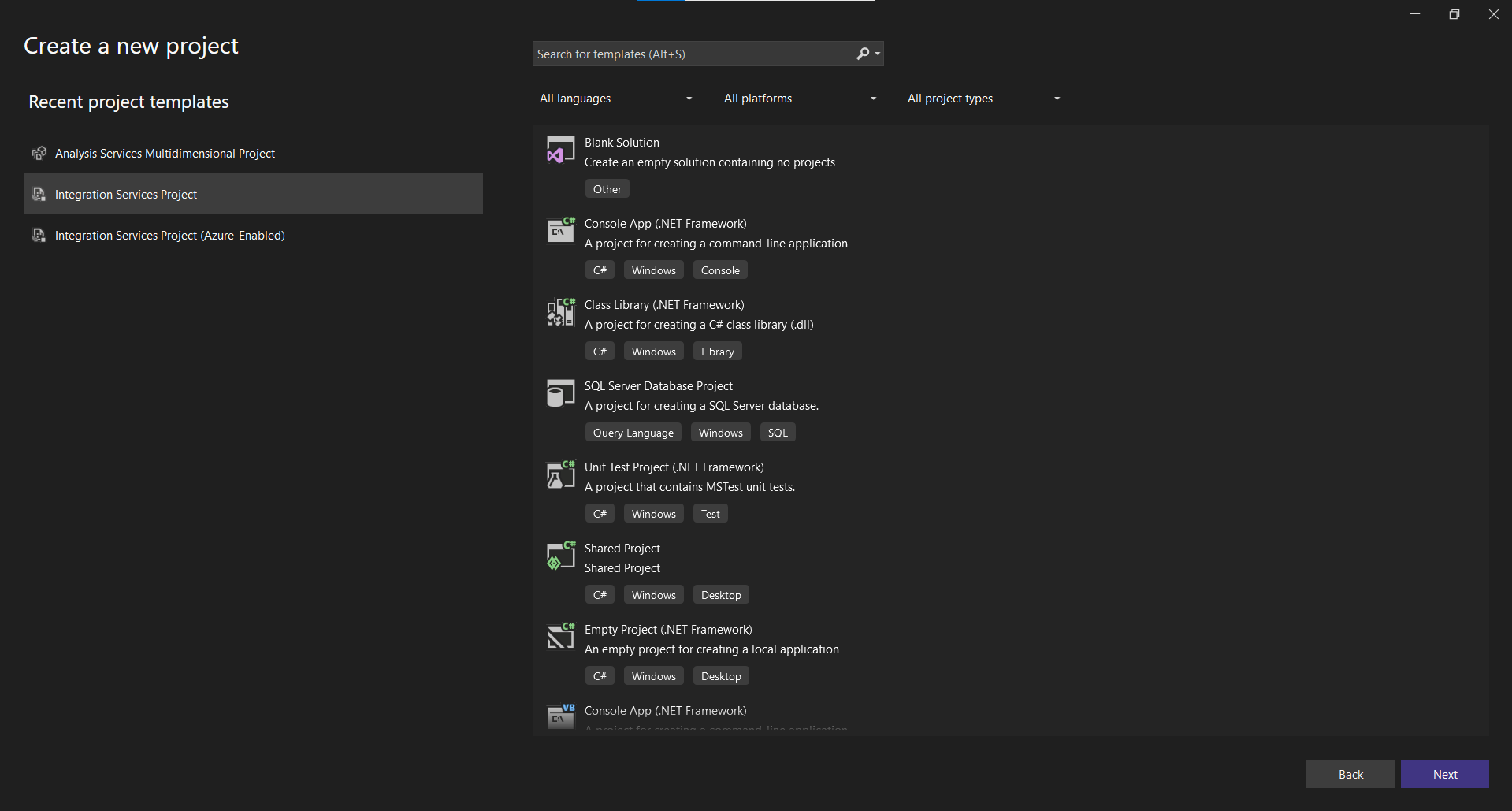
Vì lí do một số trường dữ liệu bị rỗng nên chúng em sử dụng ngôn ngữ python để chuẩn hóa dữ liệu trước khi tiến hành quá trình SSIS.

## **Kết quả**



## **Quá trình ETL**

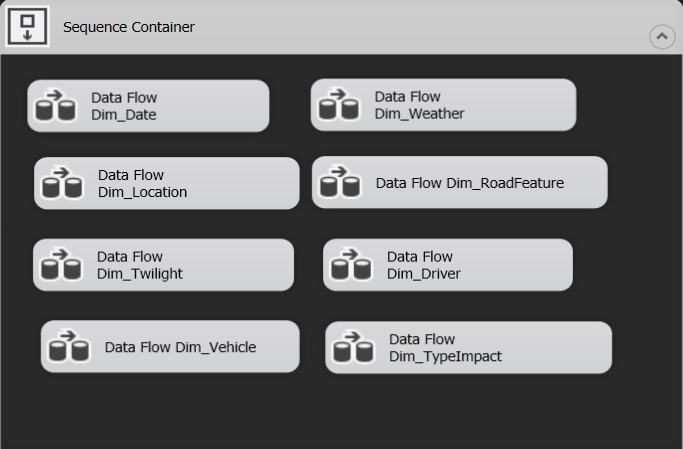
### ***Tạo dự án SSIS***



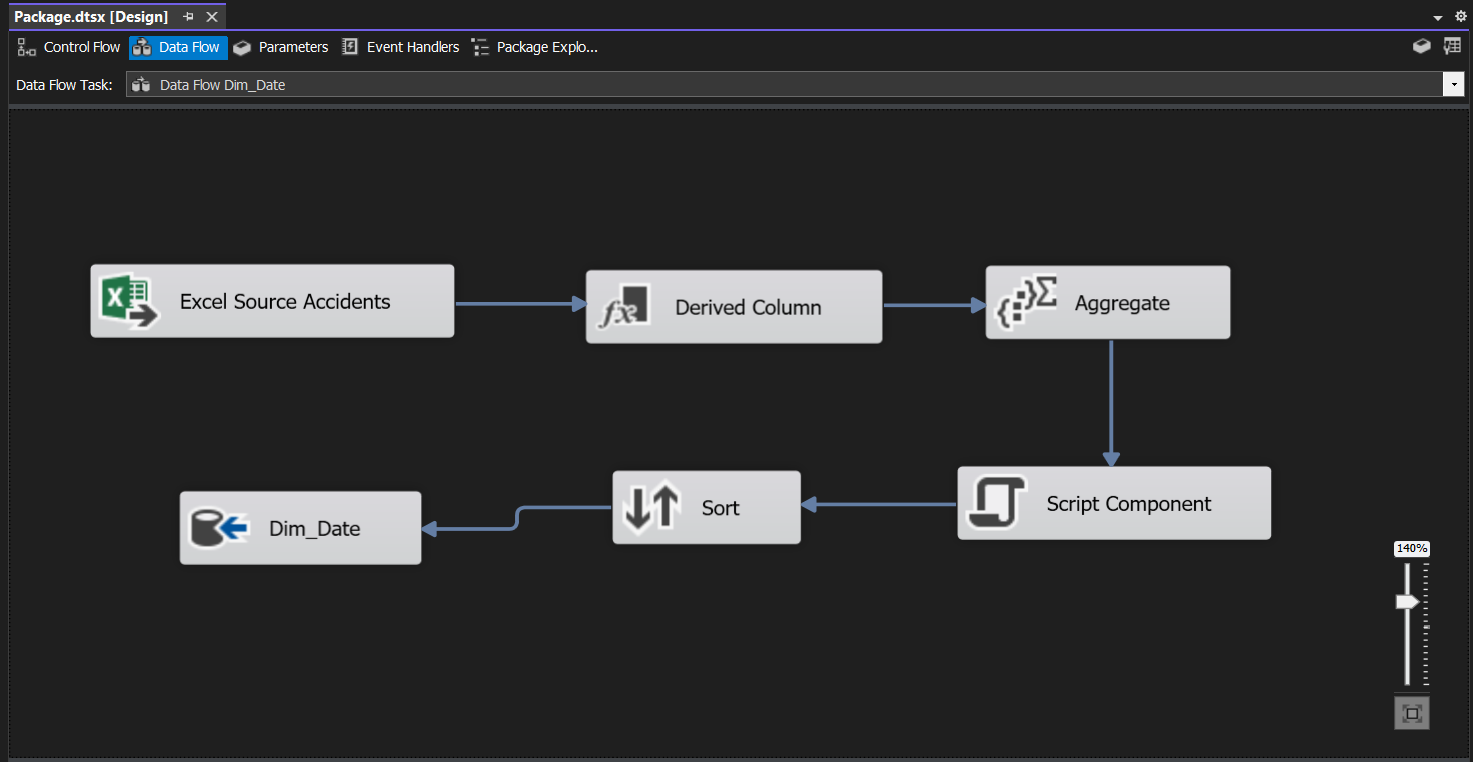
### ***Kết nối dữ liệu với sql***

### 

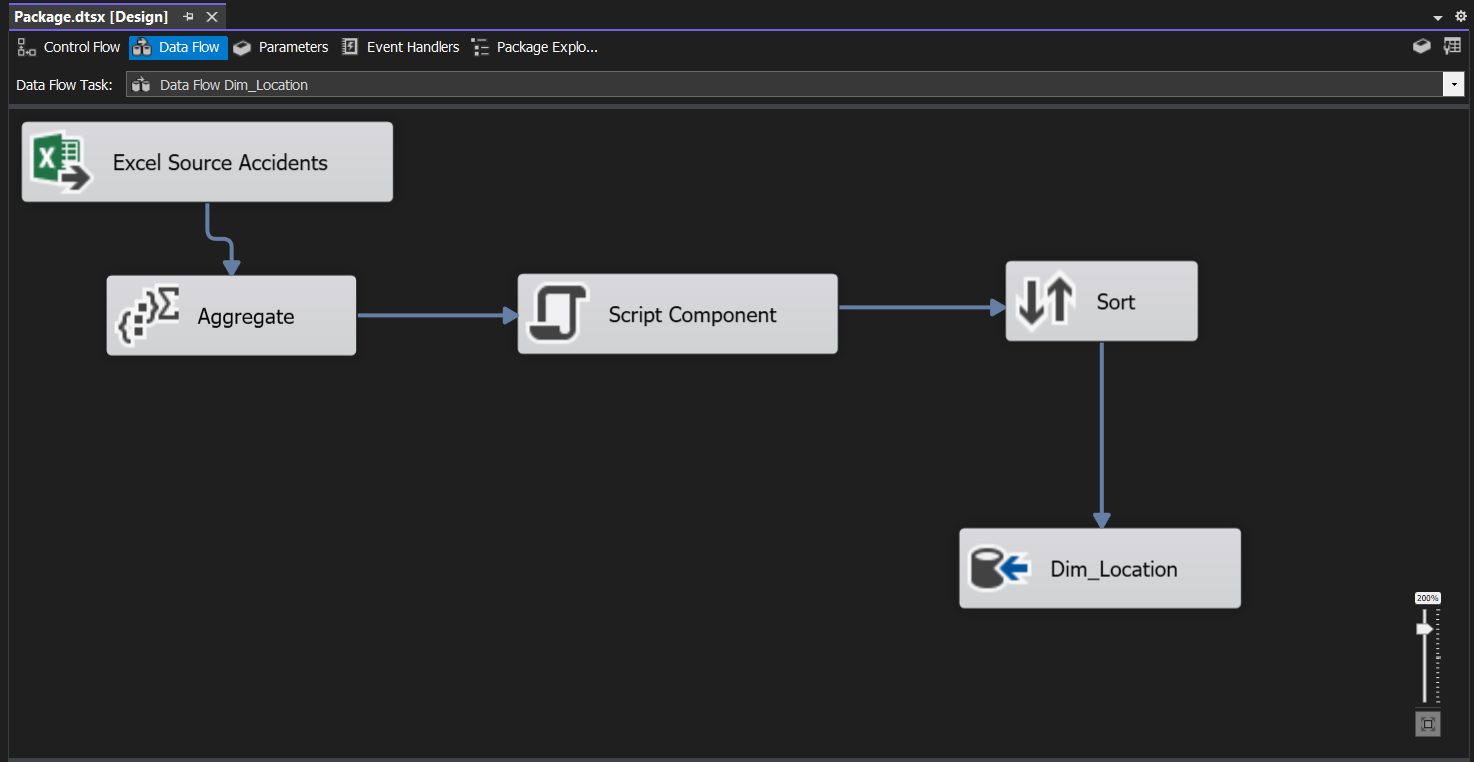
### ***Tạo các bảng Dimension***



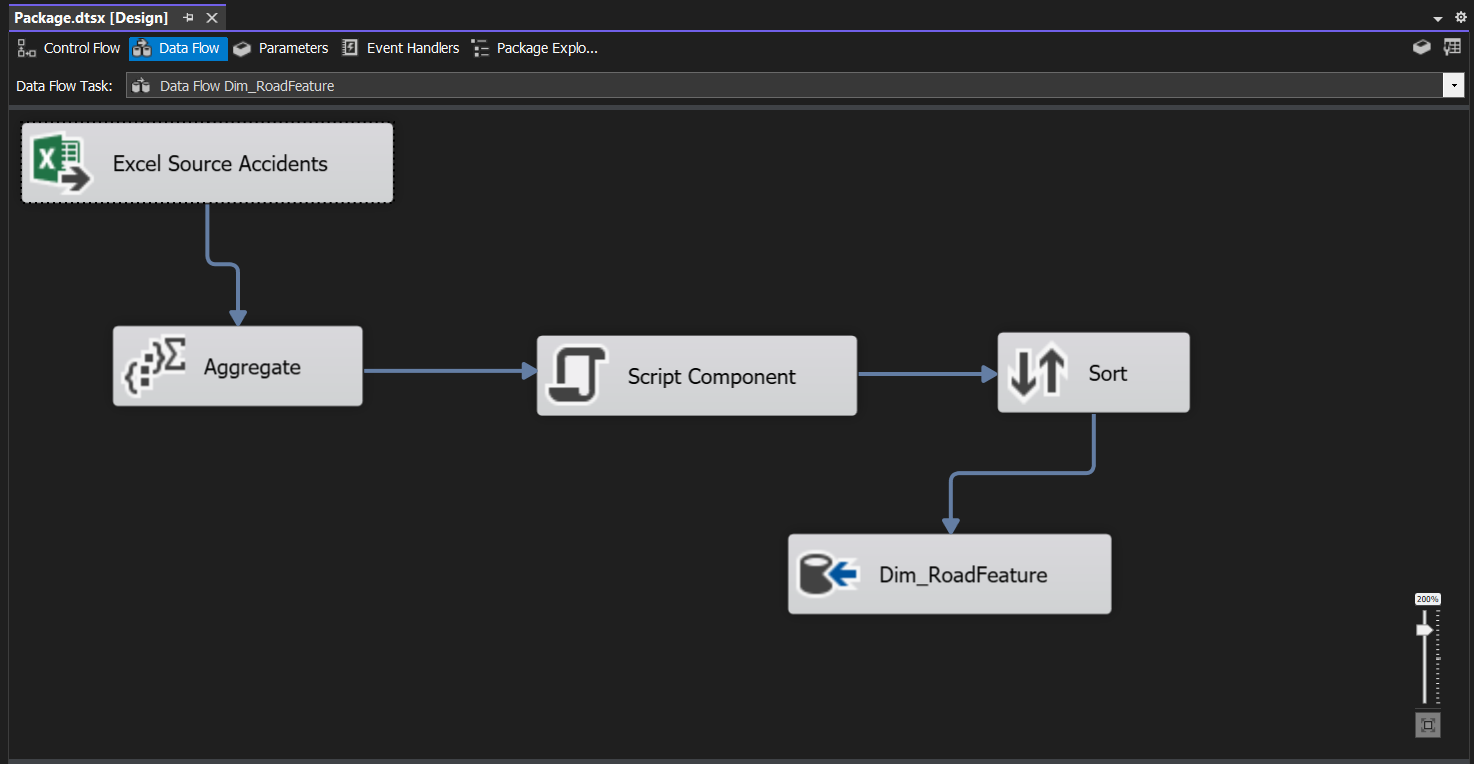
#### *Dim\_Date*



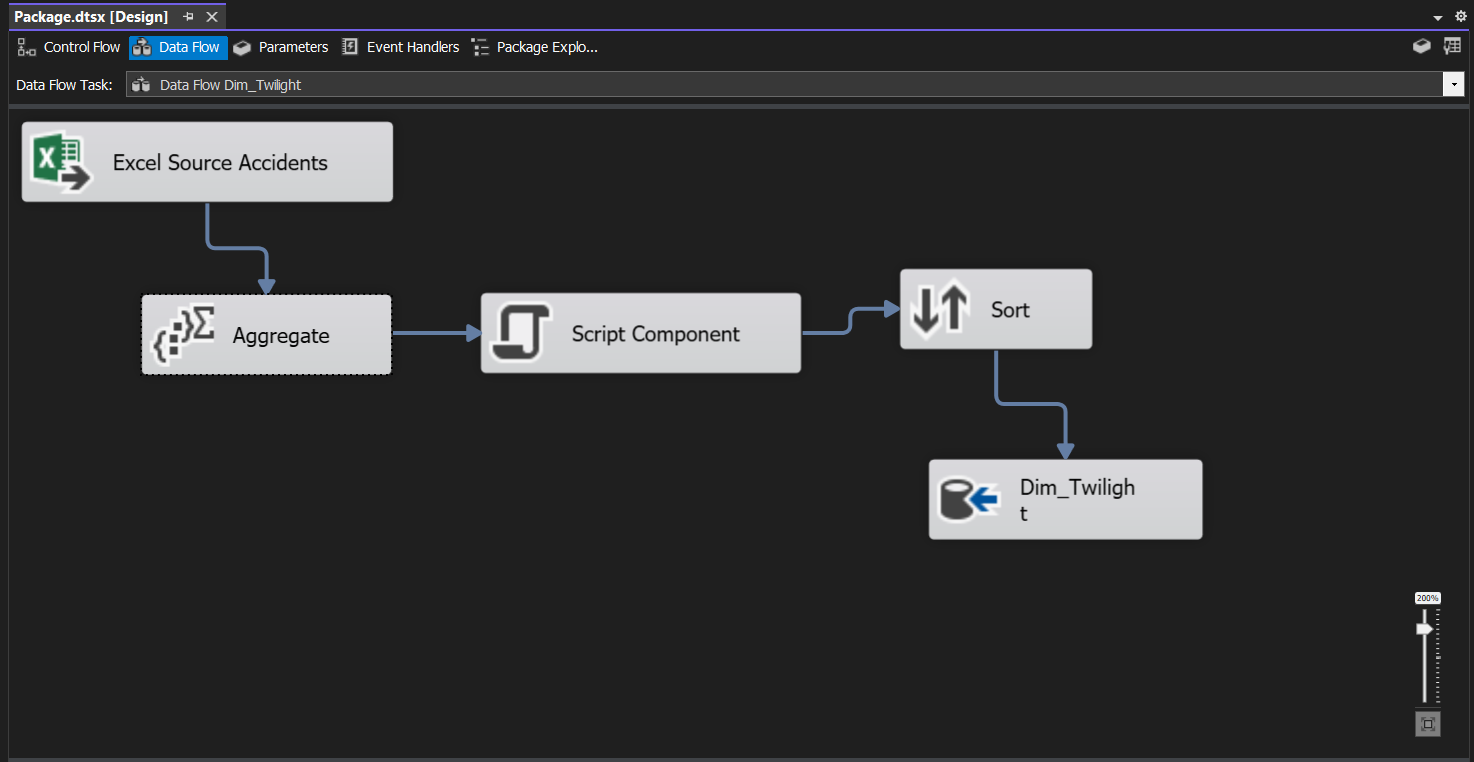
#### *Dim\_Location*



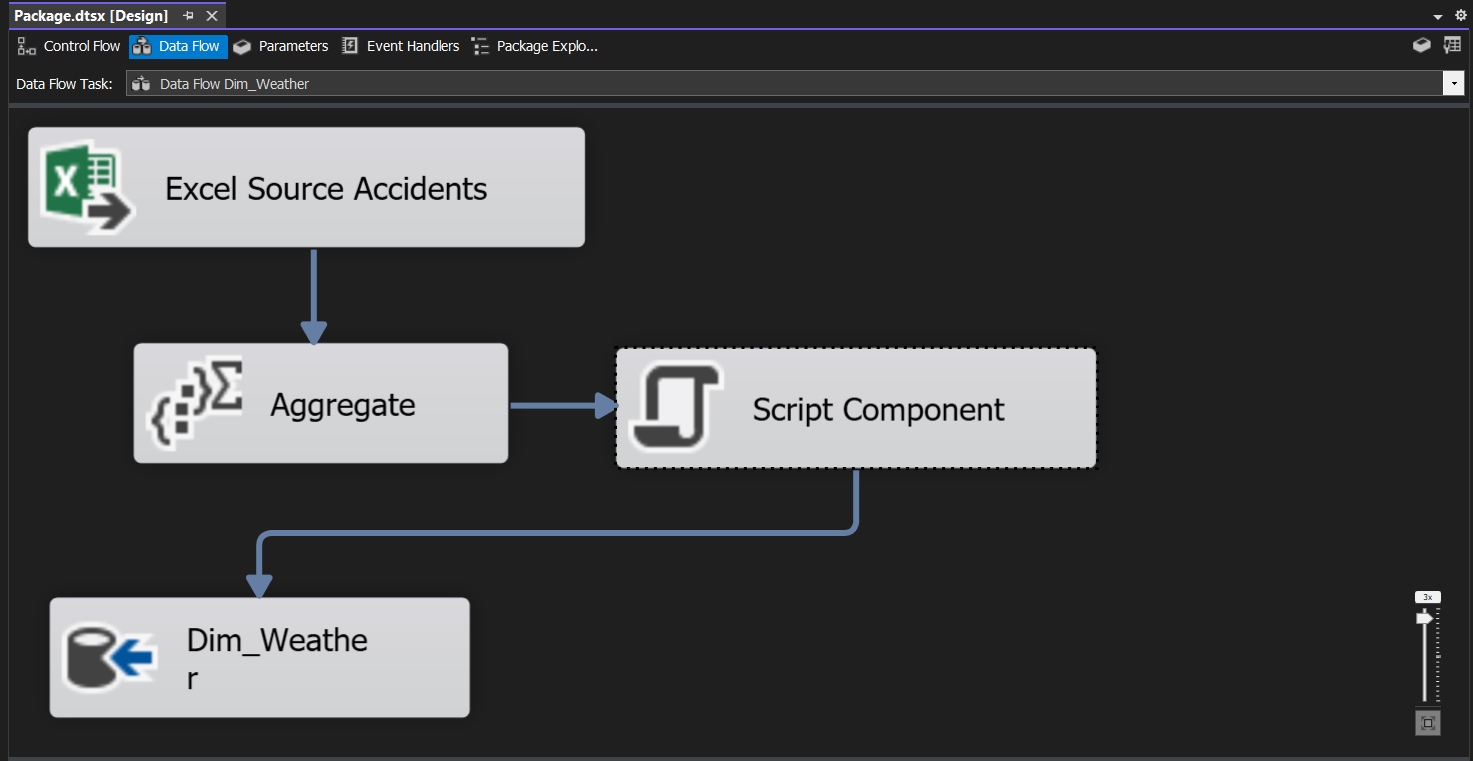
#### *Dim\_RoadFeature*



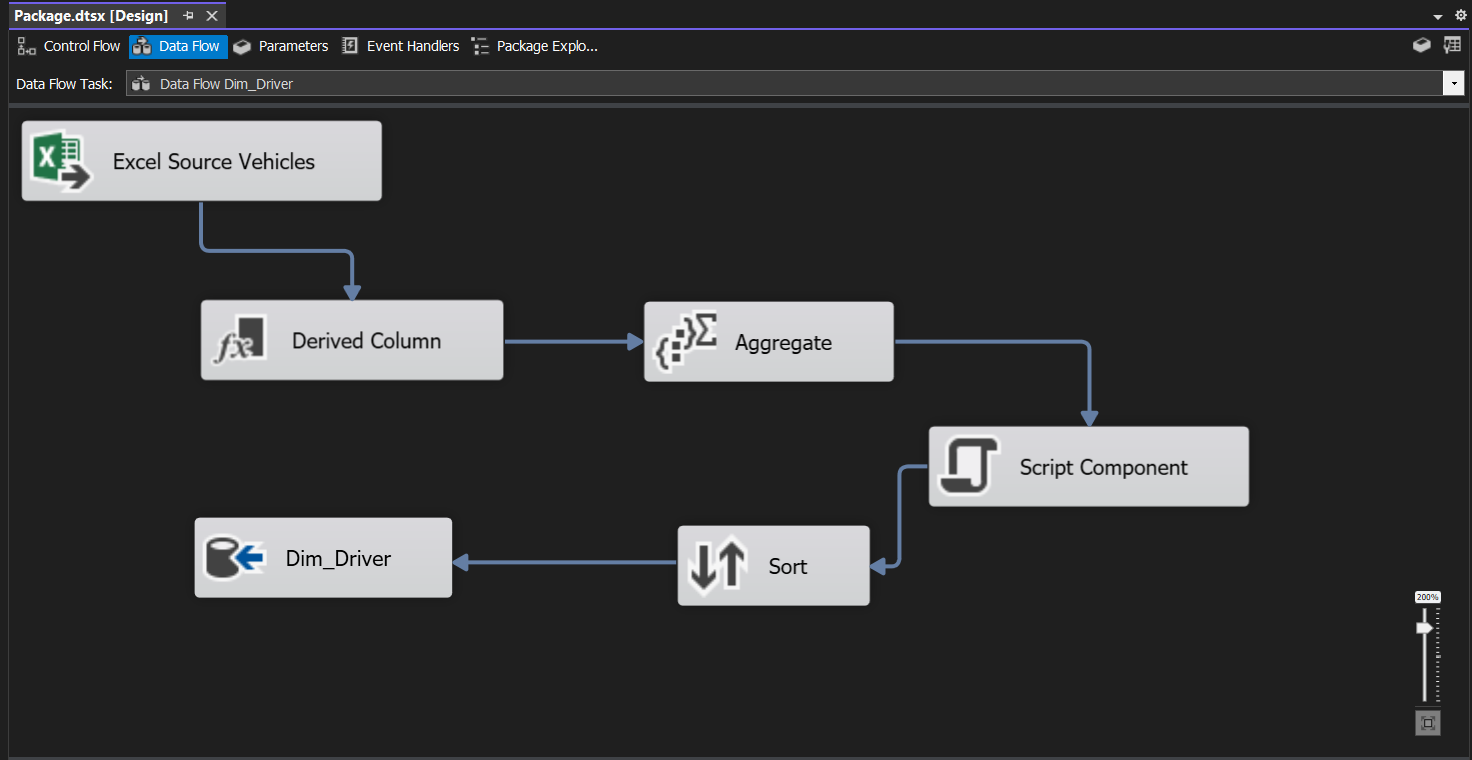
#### *Dim\_Twilight*



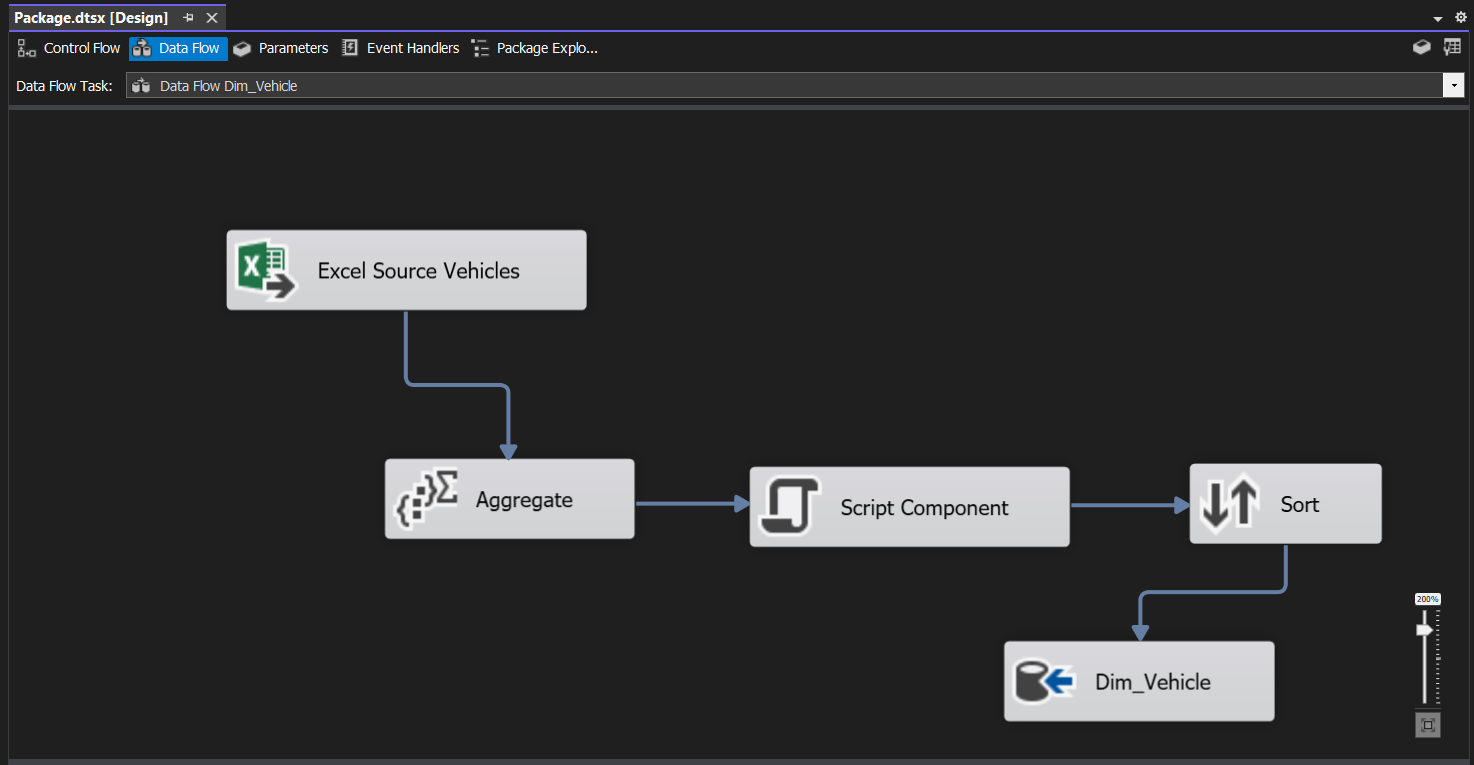
#### *Dim\_Weather*



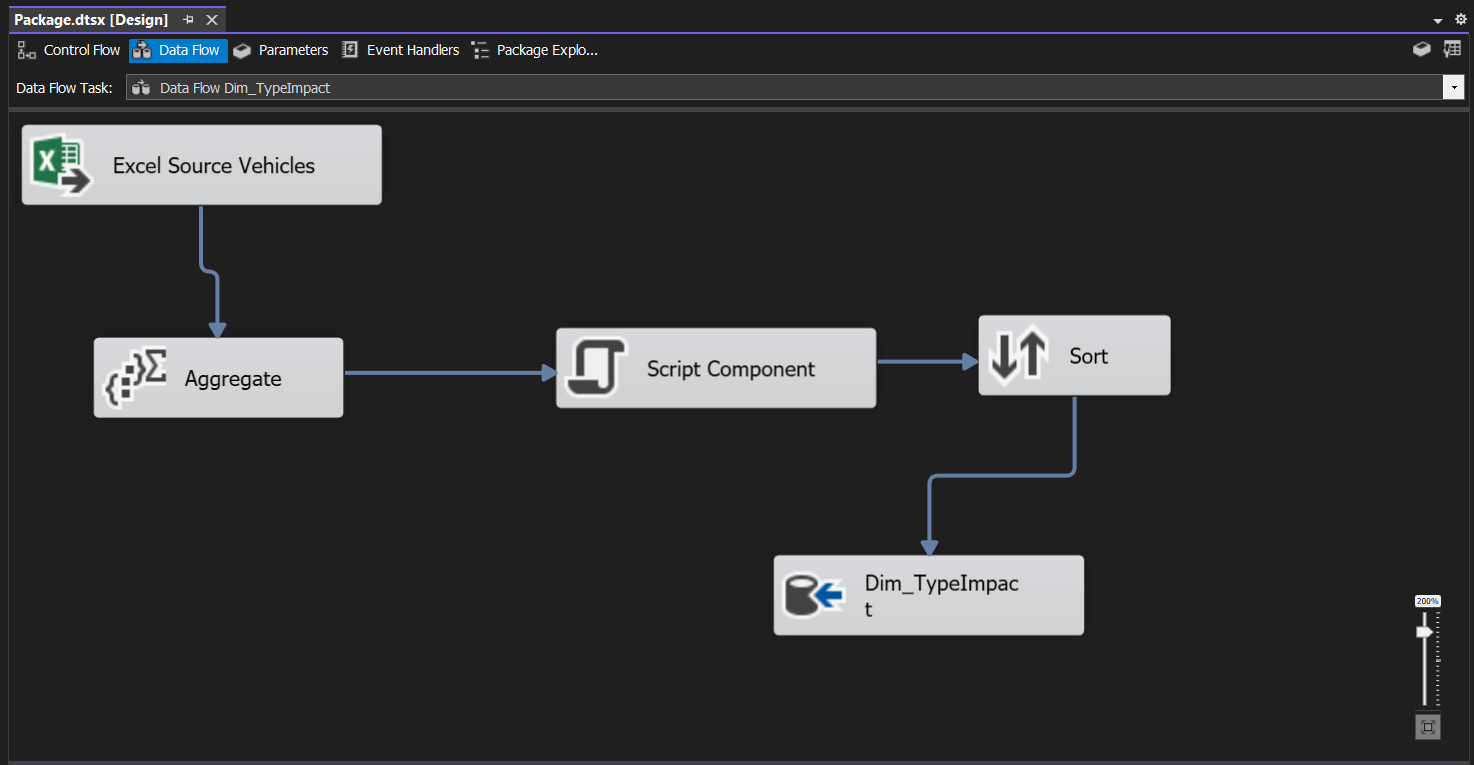
#### *Dim\_Driver*



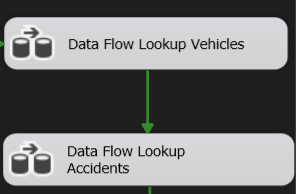
#### *Dim\_Vehicle*



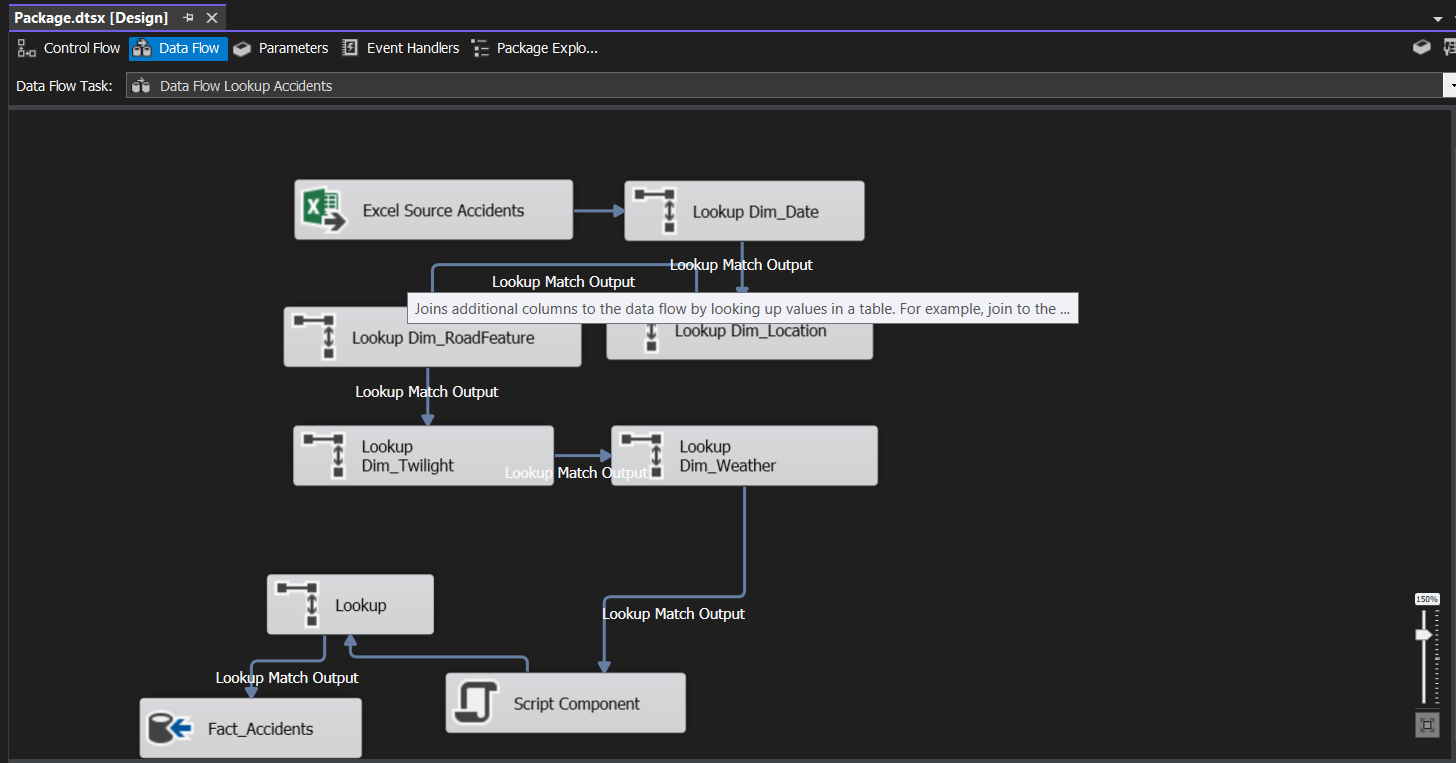
#### *Dim\_TypeImpact*



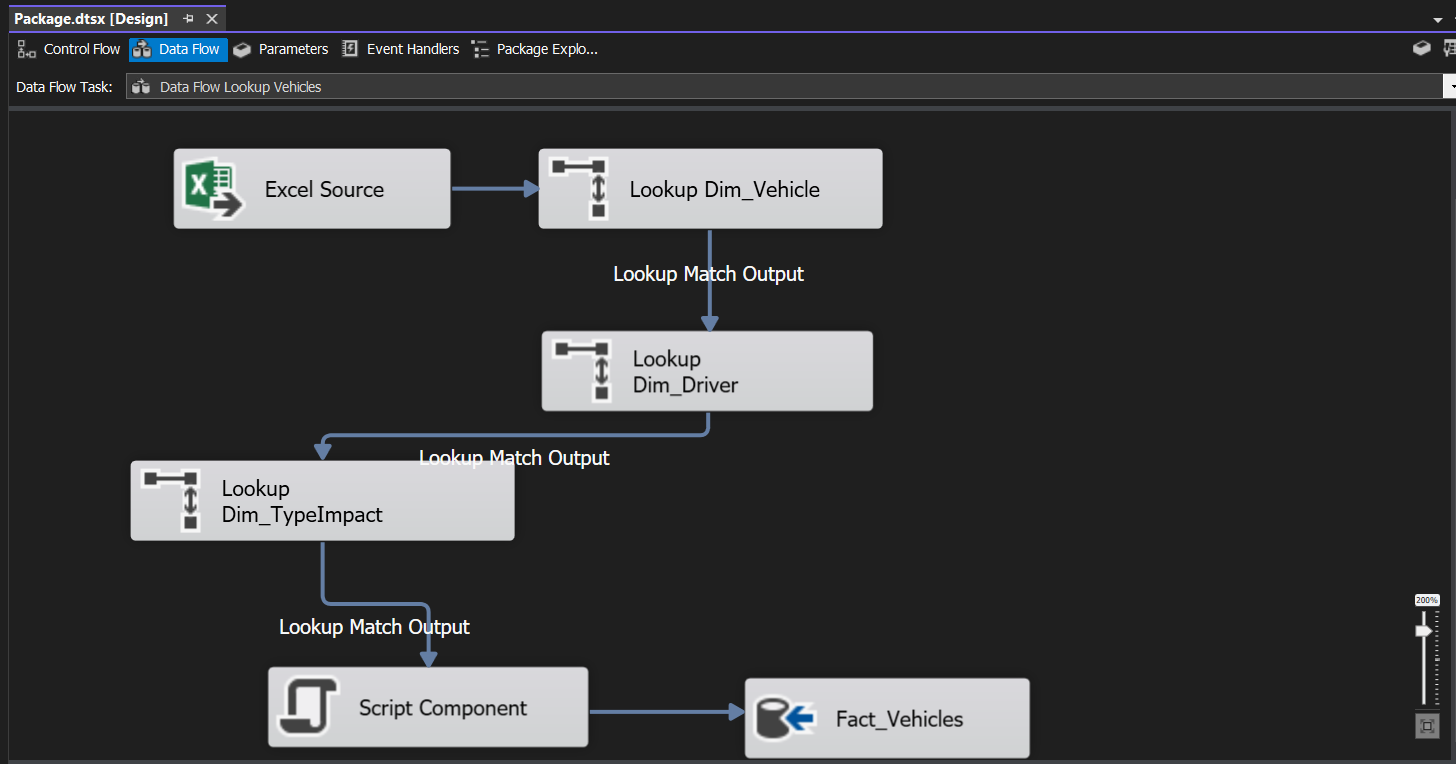
### ***Tạo bảng fact***



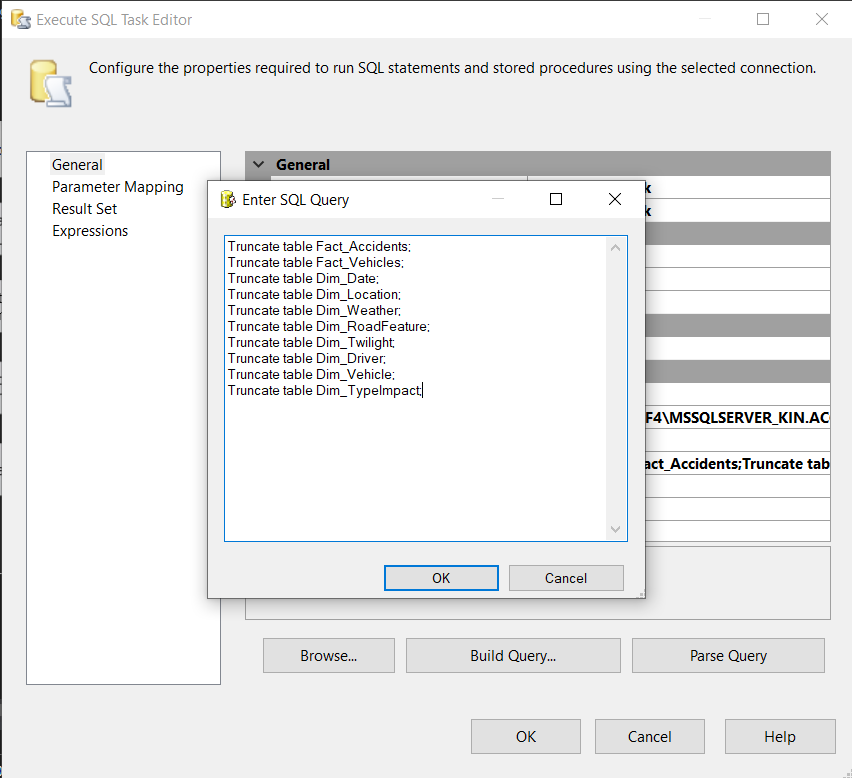
#### *Fact\_Accidents*



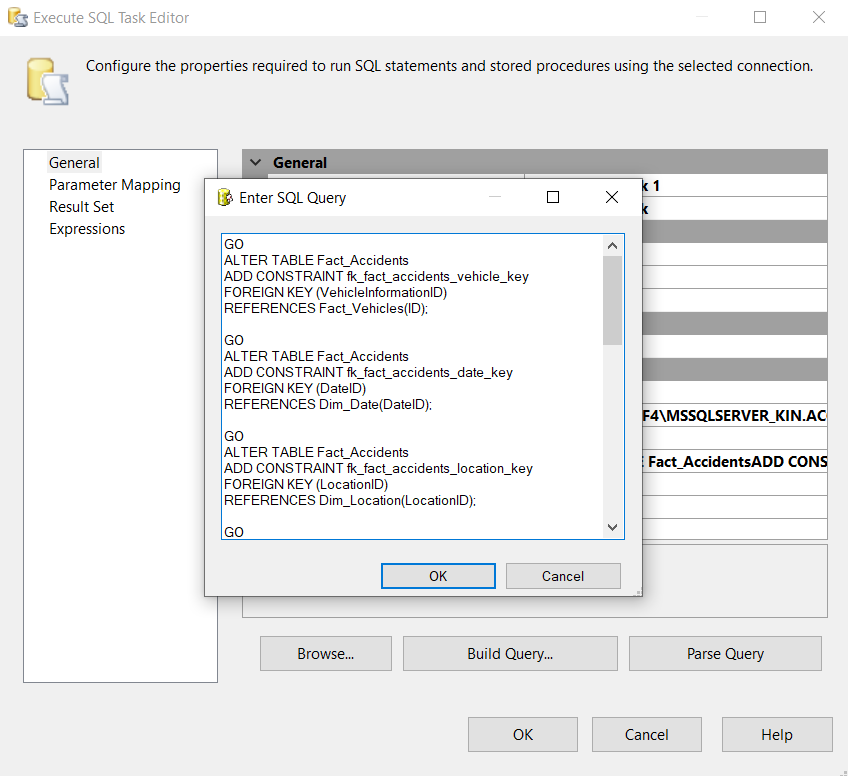
#### *Fact\_Vehicles*



### ***Xoá dữ liệu của các bảng trước khi đổ dữ liệu***



### ***Thêm Khoá ngoại giữa bảng Fact và bảng Dim***



**KẾT LUẬN**

## **KẾT QUẢ ĐẠT ĐƯỢC**

Sau quá trình xây dựng và triển khai mô hình kho dữ liệu về tai nạn giao thông tại Mỹ, chúng tôi đã đạt được một số kết quả quan trọng:

**Hệ thống kho dữ liệu hoàn chỉnh**: Một hệ thống kho dữ liệu đã được thiết lập, bao gồm các bảng chiều (Dimensions) và bảng sự kiện (Facts) để lưu trữ và quản lý dữ liệu về tai nạn giao thông.

**Phân tích và báo cáo dữ liệu**: Qua việc sử dụng kho dữ liệu, chúng tôi đã có thể thực hiện các phân tích chi tiết về tai nạn giao thông tại Mỹ, từ đó rút ra các kết luận về các yếu tố ảnh hưởng đến tai nạn như thời gian, vị trí và điều kiện đường xá. Các báo cáo phân tích được tạo ra giúp các cơ quan chức năng hiểu rõ hơn về mô hình tai nạn và đưa ra các biện pháp cải thiện an toàn giao thông.

**Dễ dàng truy xuất dữ liệu**: Hệ thống kho dữ liệu đã được tối ưu hóa để có thể truy xuất thông tin một cách nhanh chóng và hiệu quả, hỗ trợ các yêu cầu phân tích và ra quyết định trong thời gian thực.

**Khả năng mở rộng**: Kho dữ liệu được thiết kế sao cho có thể dễ dàng bổ sung dữ liệu mới, mở rộng với các bảng chiều hoặc sự kiện mới nếu cần thiết mà không ảnh hưởng đến hiệu suất của hệ thống.

## **HƯỚNG PHÁT TRIỂN**

**Mở rộng dữ liệu**: Hiện tại, kho dữ liệu chỉ tập trung vào các yếu tố cơ bản liên quan đến tai nạn giao thông như vị trí, thời gian, và điều kiện đường. Trong tương lai, hệ thống có thể được mở rộng để bao gồm các yếu tố khác như tác động của điều kiện thời tiết, hành vi lái xe, thông tin về các phương tiện tham gia giao thông, và các yếu tố môi trường khác.

**Tích hợp dữ liệu thời gian thực**: Để nâng cao khả năng ra quyết định, có thể triển khai việc tích hợp dữ liệu thời gian thực từ các hệ thống giám sát giao thông, như camera giao thông và các cảm biến thông minh. Điều này sẽ giúp phân tích các tai nạn giao thông ngay khi chúng xảy ra và cung cấp thông tin kịp thời cho các cơ quan chức năng.

**Phân tích nâng cao và học máy**: Dựa trên kho dữ liệu hiện tại, các mô hình phân tích nâng cao như học máy có thể được áp dụng để dự đoán tai nạn giao thông trong tương lai. Việc sử dụng các mô hình dự báo có thể giúp các nhà chức trách xác định các khu vực có nguy cơ cao và triển khai các biện pháp phòng ngừa hiệu quả hơn.

**Chia sẻ dữ liệu và hợp tác**: Một trong những hướng phát triển quan trọng là tạo ra các API hoặc cổng thông tin mở để chia sẻ dữ liệu giữa các cơ quan chính phủ, tổ chức nghiên cứu và cộng đồng. Việc này có thể thúc đẩy sự hợp tác giữa các bên trong việc cải thiện an toàn giao thông và giảm thiểu tai nạn.

**Tối ưu hóa hiệu suất**: Với khối lượng dữ liệu ngày càng lớn, một hướng phát triển quan trọng là tối ưu hóa hiệu suất của kho dữ liệu, đặc biệt trong các truy vấn phức tạp. Các công nghệ như phân vùng dữ liệu, chỉ số tối ưu và lưu trữ dữ liệu trong hệ thống đám mây có thể giúp cải thiện hiệu suất và khả năng mở rộng của hệ thống.

**DANH MỤC TÀI LIỆU THAM KHẢO**

1. Inmon, W. H. (2005). *Building the Data Warehouse* (4th ed.). Wiley Publishing.

Giới thiệu về khái niệm và các phương pháp xây dựng Data Warehouse, đặt nền tảng cho kiến trúc và quản lý dữ liệu trong doanh nghiệp.

1. Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). John Wiley & Sons.

Sách hướng dẫn chi tiết về mô hình hóa dữ liệu theo phương pháp Dimensional Modeling, tập trung vào thiết kế bảng fact và dimension.

1. Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. John Wiley & Sons.

Hướng dẫn thực hành về quy trình ETL, bao gồm các kỹ thuật trích xuất, làm sạch, chuyển đổi và tải dữ liệu vào Data Warehouse.

1. Golfarelli, M., & Rizzi, S. (2009). *Data Warehouse Design: Modern Principles and Methodologies*. McGraw-Hill.

Cung cấp các nguyên tắc và phương pháp thiết kế Data Warehouse hiện đại, phù hợp cho việc quản lý dữ liệu lớn và hệ thống phức tạp.

1. Malinowski, E., & Zimányi, E. (2008). *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*. Springer.

Khám phá các phương pháp thiết kế Data Warehouse tiên tiến, bao gồm ứng dụng trong phân tích không gian và thời gian.

1. Watson, H. J. (2002). *Data Warehousing in the Age of Big Data*. Morgan Kaufmann.

Thảo luận về vai trò của Data Warehouse trong bối cảnh dữ liệu lớn và cách tích hợp với các hệ thống dữ liệu hiện đại.

1. *Apache NiFi Documentation*. Apache Software Foundation.

Tài liệu kỹ thuật cho công cụ ETL Apache NiFi, cung cấp thông tin chi tiết về cách sử dụng và tối ưu hóa quy trình ETL.

1. *SQL Server Integration Services (SSIS) Documentation*. Microsoft.

Hướng dẫn sử dụng SSIS trong Microsoft SQL Server để thiết kế và triển khai quy trình ETL trong hệ thống Data Warehouse.